一种轻量级文本蕴含模型

王 伟[1]，孙成胜[1]，伍少梅[2]，张 芮[2]，康 睿[2]，李小俊[3]
(1. 中国电子科技网络信息安全有限公司，成都 610041；2. 四川大学计算机学院，成都 610065；
3. 卫士通信息产业股份有限公司，成都 610041)

摘 要：现有主流文本蕴含模型大多采用循环神经网络编码，并采用各种注意力推理机制或辅以手工提取的特征来提升蕴含关系识别准确率，由于复杂的网络结构和 RNNs 网络串行机制导致这些模型训练和推理速度较慢. 本文提出轻量级文本蕴含模型，采用自注意力编码器编码文本向量，点积注意力交互两段文本，再采用卷积神经网络对交互特征推理，整个结构可根据不同数据的推理难度叠加不同模块数量. 在多个文本蕴含数据集实验表明，本文模型在保持较高识别准确率情况下仅用一个块参数仅为 665 K，模型推理速度相比其他主流文本蕴含模型至少提升一倍.

关键词：注意力机制；卷积神经网络；轻量级；文本蕴含
中图分类号：TP391 **文献标识码**：A DOI：10.19907/j.0490-6756.2021.052001

A lightweight text entailment model

WANG Wei[1], *SUN Cheng-Sheng*[1], *WU Shao-Mei*[2], *ZHANG Rui*[2], *KANG Rui*[2], *LI Xiao-Jun*[3]
(1. China Electronic Technology Cyber Security Company Limited, Chengdu 610041, China;
2. College of Computer Science, Sichuan University, Chengdu 610065, China;
3. Westone Information Industry INC, Chengdu 610041, China)

Abstract: Most of the existing mainstream textual entailment models adopt recurrent neutral network to encode text, and various complex attention mechanisms or manually extracted text features are used to improve the accuracy of textual entailment recognition. The training and inference speed of the models is usually slow due to the complex network structure and the sequential nature of RNNs. In this paper, Lightweight Text Entailment Model is proposed. In the proposed model, the self-attentional encoder is adopted to encode text vectors; the dot product attention mechanism is adopted to interact two texts; the convolutional neural network is adopted to deduce interactive features, and the module number of the structure can be adjusted according to the reasoning difficulty of data. Experiments on multiple datasets show that the parameter size of single module in the model is only 665 K, and the inference speed of the model is at least twice as high as that of other mainstream models, under the condition of high accuracy.
Keywords: Attention mechanism; CNN; Lightweight; Textual entailment

1 引 言

文本蕴含识别(Recognizing Textual Entailment, RTE)是自然语言理解研究中一项基本任务，对问答对话，阅读理解，文本摘要等任务有辅助作用. 文本蕴含任务定义为给定一对文本，分别称

收稿日期：2021-06-28
基金项目：四川省新一代人工智能重大专项(2018GZDZX0039)；四川省重点研发项目(2019YFG0521, JG2020125)
作者简介：王伟(1975—)，男，硕士，四川成都人，主要研究领域为网络信息安全. E-mail：449685750@qq.com
通讯作者：伍少梅. E-mail：wu_scdx@126.com

为前提(Premise)和假设(Hypothesis),模型需要推理出这两段文本之间的关系,文本关系主要包括蕴含、中立和矛盾.

基于神经网络的文本蕴含模型取得较高的识别准确率.现有主流文本蕴含模型[1-2]通常使用双向 LSTM 网络编码文本,并采用注意力机制对两段文本交互,再用循环神经网络分析并推理交互特征从而判断文本关系.这些模型一个共同特点是采用多次循环神经网络编码和推理文本,并且近年来部分文本蕴含模型[3-4]为了进一步提升精度,模型整体趋势构建得越来越复杂,花费更多的参数,模型也需要花费更长的训练和推理时间.更高的训练代价使得模型难以快速微调和迭代,更长的推理时间使得模型不适于线上实际应用.

本文致力于探索高效的文本蕴含模型,采用并行架构,在保证模型精度情况下尽量精简模型结构,提高模型运行速度.本文提出轻量级文本蕴含模型(Lightweight Textual Entailment Model, LwTEM),模型采用改进的自注意力编码器分别编码前提和假设文本向量,捕捉文本长距离依赖,采用点积注意力机制深入比较两段文本语义,然后用卷积神经网络分别提取两段文本局部交互特征,模型可堆叠多个模块进一步强化推理效果,凝练高层语义信息.

2　相关工作

基于神经网络方法是文本蕴含任务的主流方法,主要分为两类.一类通过构建更好的文本编码器来独立编码两段文本,再采用神经网络分类器对两个文本向量分类.编码器主要包括循环神经网络[5],卷积神经网络[6],自注意力网络[7].一个好的文本编码器可适用于其他文本任务,但由于该方法没有两段文本的交互过程,分类器难以捕捉复杂的文本关系.另一类方法采用交互聚合框架来建模文本关系,在文本编码后采用注意力机制从语义上比较两段文本,再将文本交互特征聚合后进行推理.Chen 等人[2]采用词级别注意力矩阵对齐交互文本,并使用双向 LSTM 网络编码和聚合文本特征.

在这种框架基础上,主要有 4 种方式被用于进一步提升文本蕴含性能.(1) 是采用手工提取文本特征来作为模型输入,Chen 等[2]采用句法解析树得到句法特征,Tay 等[1]和 Gong 等[8]使用词性特征,Kim 等[4]和 Gong 等[8]手动提取字符匹配特征 (2) 是采用复杂的文本对齐过程,Wang 等[9]采用多视角匹配操作,Tan 等[10]采用多种交互对齐方式;(3) 是构建复杂的后处理过程来处理文本交互结果,Tay 等[1]使用因子分解层增强文本交互表示,Gong 等[8]采用密集连接操作构建深度卷积网络从交互结果中抽取文本信息. Xiong 等[11]使用门控机制处理句子级交互信息,结合单词级细粒度推理机制捕获全局语义;(4) 通过多次迭代推理或多层编码器来提取文本深层语义信息. Liu 等[3]使用循环神经网络多次迭代推理文本交互特征. Kim 等[4]堆叠编码层和交互对齐层,采用自动编码器处理大量特征空间. Tay 等[12]采用多种层级的注意力强化推理结果,并用密集连接加强多层级注意力传播.

这些模型大多数都采用循环神经网络进行编码和推理,许多基于循环神经网络的文本任务[13-14]取得了成功, RNN 网络的结构和串行数据处理方式,非常适用于文本数据,但该网络运行时间通常较慢,再加上近年来文本蕴含模型各种复杂的特征和组件的堆叠使得模型结构越来越复杂,参数越来越多,训练和推理时间也越来越长.

Vaswani 等[15]提出 Transformer 结构,摒弃了串行编码结构,采用自注意力和全连接网络作为基本架构,在多个文本任务上取得了较好的结果.自注意力可跨距离捕捉文本远距离依赖,良好的并行能力使得该结构在速度上很有优势,便于堆叠多层,可提取更准确的高层语义信息. Zhang 等[16]把语义角色融合到自注意力机制应用于 RTE 任务.卷积神经网络擅长于局部特征建模,也被广泛使用文本任务中. Xu 等[17]将上下文相关的概念特征整合到卷积神经网络中应用于短文本分类任务.卷积核及参数共享特点使得该网络在运行速度和参数数量上也要优于循环神经网络.

3　模　型

轻量级模型 LwTEM 包括嵌入层,编码层,交互层,特征提取层和输出层,整体架构如图 1 所示.其中编码层、交互层和特征提取层共同构成一个模块,如图 1 蓝色虚线部分,模型针对不同数据集特点可叠加多个模块,达到不同的推理效果.

3.1　嵌入层

由于编码层的自注意力网络无法捕捉文本的位置信息,因此本文在嵌入层将预训练词向量和位置编码拼接作为模型输入.位置编码采用 Vaswani 等[15]提出的相对位置编码.

$$PE_{(\mathrm{pos},2l)}=\sin(\mathrm{pos}/1000^{2l/d_{\mathrm{model}}}) \tag{1}$$

$$PE_{(\mathrm{pos},2l+1)}=\cos(\mathrm{pos}/1000^{2l/d_{\mathrm{model}}}) \tag{2}$$

式中，pos 表示文本序列中当前词的位置；l 表示位置向量中第 l 维；sin 和 cos 分别表示奇数维度和偶数维度用的数学函数. 该位置函数可以根据 sin 和 cos 函数的数学特性捕捉单词之间的相对位置关系.

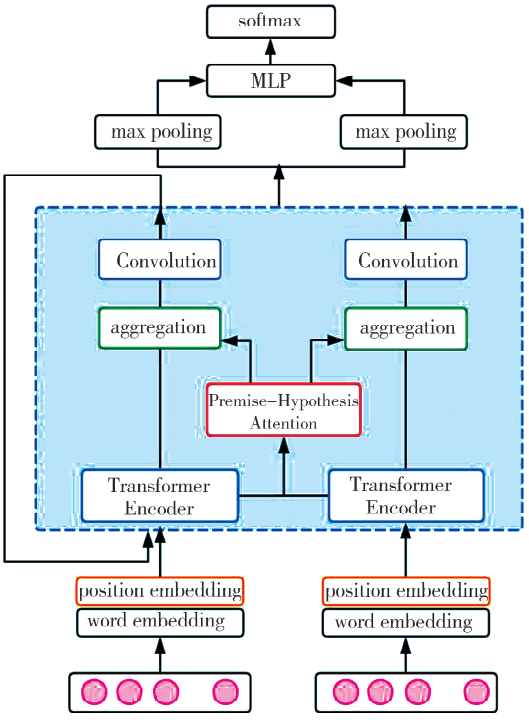


图 1　LwTEM 模型架构

Fig. 1　Model architecture of LwTEM

3.2　编码层

本文编码层和 Transformer 编码器结构相似，由两个子模块组成：多头注意力和两层前馈神经网络，每个子模块通过残差和层归一化相连.

自注意力可无视距离交互远距离文本，但同时也会忽略文本序列信息. 对于一句话，中心词附近的文本词的作用应该是高于远距离文本的，比如："The man in a black shirt is standing next to black box"，句子中第一个"black"和第二个"black"对于"man"这个中心词来说重要程度是不一样的，而原始自注意力计算方式将这两个"black"同等对待. 本文模型对自注意力结果增加参数约束，用于学习不同距离的词对中心词产生的不同效果. 虽然嵌入层的位置编码也可缓解该问题，但实验显示增加参数约束后效果更好. 改进后自注意力计算方式如图 2 所示.

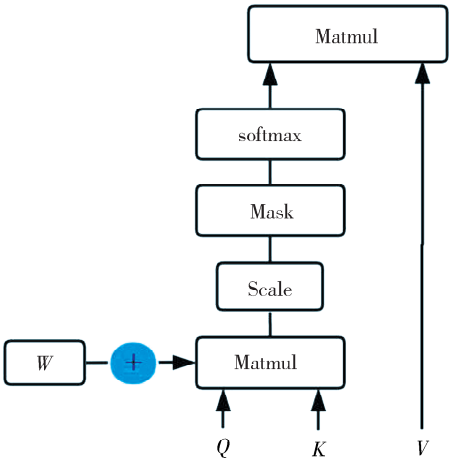


图 2　模型自注意力方式

Fig. 2　Structure of self attention

计算方式如式(3)所示。

$$Attention(Q,K,V)=\mathrm{softmax}\left(\frac{QK^T+W}{\sqrt{d_k}}\right) \tag{3}$$

式中，Q,K,V 为输入文本分别采用不同参数的线性映射得到；d_k 为多头自注意力的隐层. 本文模型在原本的自注意力计算方式下增加参数 W，W 和自注意力矩阵大小相同，为可训练参数，用于改善不同距离词的自注意力效果.

多头自注意力计算结果通过拼接送入两层前馈神经网络中，如式(4)所示.

$$F(p)=W_2(\mathrm{Relu}(W_1p+b_1))+b_2 \tag{4}$$

式中，W_1、W_2 和 b_1、b_2 分别为两层的前馈神经网络的权重和偏置；p 为经过多头自注意力计算的前提文本向量，假设文本向量计算方式也同上. 前馈神经网络将自注意力编码后的信息映射到高维空间，通过非线性激活函数 Relu 进一步选择有效的特征.

3.3　交互层

模型的交互方式采用简单高效的点积注意力，得到两段文本的词级相似度结果.

$$\mathrm{Inter}(p_i,h_j)=p_i{}^T\cdot h_j \tag{5}$$

$$\tilde{h}_j=\sum_i\mathrm{softmax}(\mathrm{Inter}(h_j,p_i))p_i \tag{6}$$

$$\tilde{p}_i=\sum_j\mathrm{softmax}(\mathrm{Inter}(p_i,h_j))h_j \tag{7}$$

式中，p_i 表示编码后的前提文本第 i 个词；h_j 表示假设文本第 j 个词，点积交互结果通过 softmax 函数归一化. 交互向量 $\tilde{p}_i$ 为假设文本每个词和前提文本第 i 个词的注意力权重相乘相加得到的结果，表示第 i 个前提词相关的假设文本向量，$\tilde{h}_j$ 由前提文本每个词和 h_j 做点积操作得到交互权重，再分别和前提文本相乘相加得到. 这种计算方式可得到

和另一个文本序列相关的交互特征向量.

将交互特征向量和原编码向量通过以下方式进行融合.

$$\overline{p}_i^1=[p_i;\widetilde{p}_i] \tag{8}$$

$$\overline{p}_i^2=[p_i;\widetilde{p}_i\circ p_i] \tag{9}$$

$$\overline{p}_i^3=[p_i;\widetilde{p}_i-p_i] \tag{10}$$

$$\overline{p}_i=[\overline{p}_i^1;\overline{p}_i^2;\overline{p}_i^3] \tag{11}$$

式中,[;]表示连接符号;∘表示元素相乘,可表示原文本向量和交互向量之间的相似度,相减操作可表示找到向量之间的差异. 假设文本的特征融合方式也同上.

3.4　特征提取层

特征提取层采用卷积神经网络提取融合的特征. 由于编码层的自注意力网络可充分编码远距离文本信息,而卷积神经网络更关注局部特征,结合 CNN 网络的局部感知和自注意力编码的全局信息,可全面获取文本高层语义和蕴含特征. 卷积计算方式如式(12)所示.

$$\hat{p}_i=f(w\cdot\overline{p}_{i,i+h-1}+b) \tag{12}$$

其中,b 是偏置项;f 是带 Relu 激活函数的非线性映射;w 为卷积核;h 为滑动窗口大小,过滤器和窗口内的词进行卷积操作捕捉文本局部短语信息,得到新的前提文本向量 $\hat{p}_i$. 假设向量计算方法同上.

3.5　输出层

输出层采用最大池化操作将文本特征转为固定长度的向量. 最后将前提和假设文本向量连接,采用两层前馈神经网络连接用于最后关系分类.

$$\hat{y}=\mathrm{softmax}(f_2(\mathrm{Relu}(f_1([\hat{p},\hat{h},\hat{p}-\hat{h},\hat{p}*\hat{h}])))) \tag{13}$$

$$f_k=xW_k+b_k,k\in\{1,2\} \tag{14}$$

式中,$\hat{p},\hat{h}$ 为模型最终输出的前提和假设文本向量;W_k 和 b_k 为不同层前馈神经网络参数;Relu 为激活函数;$\hat{y}$ 为模型最终预测类别. 模型采用交叉熵损失训练.

4　实验结果

4.1　数据集

本文 LwTEM 模型采用 SNLI, SCITAIL, MultiNLI 三个数据集进行验证. 其中 SNLI 数据集[18]是由斯坦福在 2015 年发布的大型文本蕴含数据集,SCITAIL 数据集[19]是根据科学类多选问答任务构造的科学类文本蕴含数据集. MultiNLI 数据集[20]为 SNLI 扩展的文本蕴含数据集,在语料范围覆盖度和推理难度上都有一定加强. 表 1 展示了三个实验数据集详细信息.

表 1　3 个实验数据集分布

Tab. 1　Distribution of three experiment datasets

数据集	大小		标签
SNLI	train	549,367	entailment neutral contradiction
	dev	9,842	
	test	9,824	
MultiNLI	train	392,703	entailment neutral contradiction
	dev	20,000	
	test	20,000	
SCITAIL	train	23,596	entailment neutral
	dev	1,304	
	test	2126	

表 1 中,Entailment 表示蕴含;neutral 表示中立;contradiction 表示矛盾类别;文本蕴含评价指标为准确率(Accuracy,ACC).

4.2　实验结果

本文的模型是基于 Tensorflow 框架搭建,采用 ADAM 优化器[21]作为整个模型的优化函数. SCITAIL 数据集的学习率为 0.001,batch size 大小为 32,隐层维度为 256,SNLI 和 MultiNLI 数据集的学习率采用 warm up 策略,初始学习率为 0.000 1,分别逐步上升到 0.001 和 0.002 之后开始下降,batch size 为 256,网络隐层维度为 200. 为防止过拟合,采用 dropout[22]比率为 0.2,CNN 网络过滤器大小为 3. 输入为 300 维的 Glove 词嵌入[23]和 50 维的位置向量,对于词表外的单词,采用高斯分布随机初始化一个 300 维的向量,所有词向量在整个训练过程中不更新.

表 2 至表 4 分别显示了三个数据集的准确率结果,由于仅 SNLI 数据集中部分对比模型列出了模型参数数量,因此仅在 SNLI 数据集上对比模型参数数量结果,表 5 对比了部分模型的推理速度.

表 2 为 SNLI 数据集的结果,LwTEM 模型在 SNLI 测试集上达到了 88.4%的准确率,模型参数数量却只有 665 K,远低于其他主流模型参数量.

BiMAP 网络[9]采用全匹配,最大匹配,平均匹配和注意力匹配多种对齐过程丰富文本匹配特征,CAFE 网络[1],DIIN 网络[8]在输入层构建多种文本输入特征,SAN 网络[3]采用多次迭代推理凝练深层语义信息. 这些网络除了 DIIN 模型外都是采用循环神经网络用于文本编码和推理,其中,SAN

模型准确率高于本文模型 0.1%，但参数数量却远多于本文模型. DecompAtt[24] 采用自注意力和全连接网络为模型主要框架，没有复杂的推理过程和额外的特征，其参数数量略少于本文模型，但本文模型准确率高于该模型 1.6%. 通过分析可知，在综合考虑准确率和参数数量指标后，LwTEM 模型性能要优于主流文本蕴含模型.

表 2　SNLI 数据集结果

Tab. 2　Result of SNLI dataset

模型	参数	训练集/%	测试集/%
DecompAtt	580K	90.5	86.8
BiMPM	1.6M	90.9	87.5
DIIN	4.4M	91.2	88.0
ESIM	4.3M	92.6	88.0
CAFE	3.5M	89.2	88.3
SAN	3.5M	93.3	88.5
LwTEM	665K	93.4	88.4

表 3　SCITAIL 数据集结果

Tab. 3　Result of SCITAIL dataset

模型	测试集/%
ESIM	70.6
DecompAtt	72.3
DGEM	77.3
HCRN	80.0
CAFÉ	83.3
RE2	86.0
CSRAN	86.7
LwTEM	87.8

表 3 为 SCITAIL 数据集实验结果，由表 2 和 3 结合可知，LwTEM 模型参数数量远少于上述大多数模型，并且在 SCITAIL 数据集上准确率也达到最好的效果.

表 4　MultiNLI 数据集结果

Tab. 4　Result of multiNLI dataset

模型	Matched/%	Mismatched /%
ESIM	76.8	75.8
MwAN	78.5	77.7
DIIN	78.8	77.8
CAFÉ	78.7	77.9
DRCN	79.1	78.4
LwTEM	78.7	78.1

表 4 显示了在 MultiNLI 数据集的实验结果. MultiNLI 测试集分为 Matched 和 Mismatched 两个部分. LwTEM 模型在 Matched 数据部分达到了 78.7%的准确率，分别高于 ESIM 模型 1.9%，MwAN 模型 0.2%，并且和 DIIN，CAFE 模型结果相当，在 Mismatch 数据上准确率为 78.1%，优于表 4 中大部分模型结果.

表 5　模型推理速度对比

Tab. 5　Compare of model inference speed

模型	推理时间/(s/batch)
RE2	0.03
BiMAP	0.05
CAFE	0.07
DIIN	0.85
CSRAN	0.28
LwTEM (1 block)	0.015
LwTEM (2 block)	0.025
LwTEM (3 block)	0.035

表 5 显示了模型不同 block 的推理速度结果，推理过程中一个 batch 标准大小为 8，最大句子长度设为 20，Yang 等[25] 模型中设定完全相同，模型运行的处理器也相同，均为 Inter Core i7，唯一的不同在于操作系统，本文模型运行操作系统为 Windows10，表中其他模型运行系统为 MacOS. CAFÉ[1]，BiMAP[9]，CSRAN[12] 模型均是由单层或多层 LSTM 网络作为文本编码，并采用各种注意力匹配推理文本关系. RE2[25] 结合 CNN 网络和注意力以及残差连接构建模型，DIIN[8] 采用 CNN 网络，空间注意力以及密集连接构建的文本蕴含模型，但由于其复杂的交互特征处理过程，使得模型速度较慢. 由表 5 中可知，本文模型 1 个模块的推理时间仅为 0.015 s，优于其他所有模型推理速度，叠加到 2 个模块的模型推理时间为 0.025 s，仍然好于其他模型，在叠加到 3 个模块后，模型推理速度为 0.035 s. 表 5 推理结果表明了本文模型 LwTEM 在推理速度上要比其他主流文本蕴含模型至少快 1 倍以上.

4.3　模型性能分析

4.3.1　结构消融分析　图 3 为 LwTEM 模型在 SCITAIL 验证集上模型结构消融结果，直观反应了不同模块的结果变化[26-27]. 由图 3 可知，本文轻量级模型(a)在 SCITAIL 验证集上最好效果为 89.4%，在(b)去掉编码层注意力中参数 w 后准确率下降了 0.3%，表明了改善后的自注意力优于原自注意力效果；在(c)去掉卷积结构后模型准确率

大幅降低了 2%,表明了在本文模型中 CNN 网络提取文本局部特征对文本对关系判断的重要性;在(d)去掉原特征融合方式而采用简单的拼接方式代替原本特征融合后,模型准确率也降低了 1.5%.在这三个不同模块中,其作用最大的是 CNN 模块,其次是特征融合方式.图 3 表明了本章模型中各个模块的有效性和不可替代性.

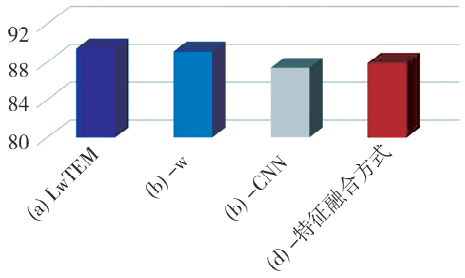


图 3　模型结构消融结果柱状图

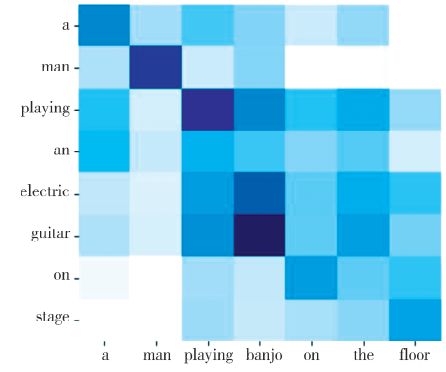
Fig. 3　The histogram of model structure ablation

4.3.2　不同 block 实验结果　表 6 为 LwTEM 模型在叠加不同模块后在各个数据集的验证集上的准确率.在 SNLI 验证集上叠加 2 个模块比 1 个模块准确率提升了 0.2%,当继续叠加到 3 个模块后,并没有发现明显的效果提升,说明本章模型在 SNLI 数据集上用 2 个模块堆叠的即可.而在 MultiNLI 数据集上叠加两个模块后,模型结果均提升了 0.6%,而继续增加到 3 个模块,准确率进一步提升 0.7%和 0.4%.受硬件限制和模型复杂度影响,本文仅测试到三个模块.由表 6 结果可知,针对不同数据集的特点,叠加不同的模块有不同的效果.对于 MultiNLI 数据集,文本对较长,句式复杂,叠加多个模块模拟多次推理效果更好,而对于简单句的 SNLI 数据集,本章模型的 1 个模块即可达到较高的文本关系识别率.

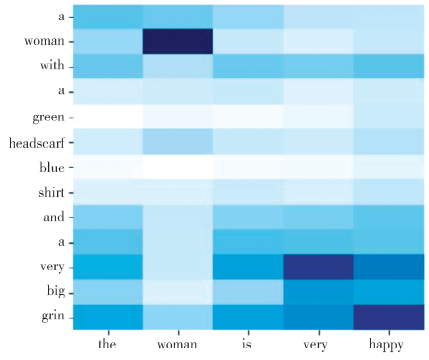
表 6　模块不同 block 结果

Tab. 6　Result of different blocks

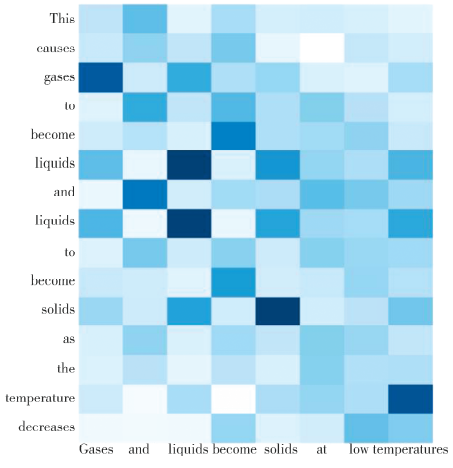
block	SNLI/%	MultiNLI/%
1 block	88.4	77.5/77.4
2 blocks	88.6	78.1/78.0
3 blocks	88.6	78.8/78.4



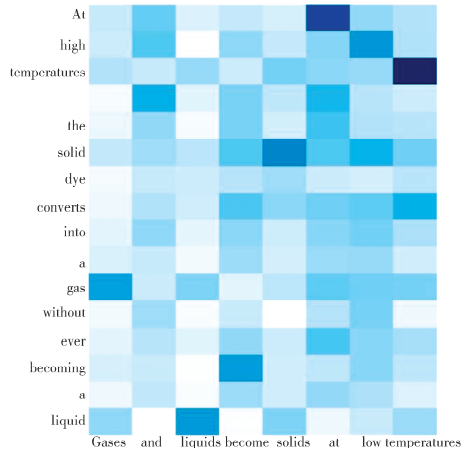
(a) 样例 1



(b) 样例 2



(c) 样例 3



(d) 样例 4

图 4　测试集样例注意力可视化图

Fig. 4　The attention visual of test cases

4.3.3　注意力可视化和案例分析　本节通过可视化注意力结果分析模型内部学到的信息. LwTEM 模型注意力模块得到的注意力热力图如 4 所示，纵轴表示前提文本，横轴表示假设文本，每个方块表示前提和假设文本中词两两对应的注意力可视化结果，颜色越深表示注意力值越大.

图 4 共包含 4 个样例，样例 1 为 SNLI 数据集中矛盾样例，样例 2 为 SNLI 数据集中蕴含样例，样例 3 为 SCITAIL 数据集的蕴含样例，样例 4 为 SCITAIL 数据集的中立样例. 以下为 4 个样例详细分析结果.

前提 1：a man playing an electric guitar on stage.

假设 1：a man playing banjo on the floor.

标签 1：矛盾

在样例 1 中，模型可准确找到“guitar”和“banjo”这一对关键词，在主语“a man”，动词“playing”一一对应的情况下，模型定位到“guitar”和“banjo”的语义差异，从而识别出文本的矛盾关系.

前提 2：a woman with a green headscarf blue shirt and a very big grin.

假设 2：the women is very happy.

标签 2：蕴含

对于样例 2 文本对，模型注意力对齐了“very big grin”和“very happy”这一对关键词，可以推测出前提蕴含假设文本语义，从而找到文本对的蕴含关系.

前提 3：This causes gases to become liquids and liquids to become solids as the temperature decreases

前提 4：At high temperatures, the solid dye converts into a gas without ever becoming a liquid

假设 3&4：Gases and liquids become solids at low temperatures

样例 3 和样例 4 的假设文本相同. 前提 3 和假设为蕴含关系，前提 4 和假设为中立关系. 通过注意力可视化分析，在样例 3 中，“gases”“become”liquids”和“liquids”“become”“solids”均有一一对应关系，“decrease”和“low”也可通过注意力推理出词义相同，在句子中谓语，宾语和多个不同主语均能一一对应情况下，可推断出文本为蕴含关系. 对于样例 4，为非蕴含关系，却有很多相同词，模型可对齐这些相同词的注意力，但是假设中关键词“liquids”到“solids”的关系，在前提中没有与之相应的关系，所以模型将其判断为非蕴含关系.

5　结　论

本文构建轻量级文本蕴含模型 LwTEM，主要由自注意力编码，注意力交互，特征融合和 CNN 网络构成，堆叠多个模块可根据不同数据集特点得到相应的推理效果. 本文在三个文本蕴含数据集上均做了实验，结果表明 LwTEM 模型在准确率和现主流模型相当情况下，参数数量和推理速度明显优于其他模型. 未来可尝试将 LwTEM 模型应用于其他文本匹配任务.

参考文献：

[1] Tay Y, Tuan L A, Hui S C. Compare, compress and propagate: Enhancing neural architectures with alignment factorization for natural language inference[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: ACL, 2018.

[2] Chen Q Zhu X D, Ling Z H, *et al*. Enhanced LSTM for natural language inference [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: ACL, 2017.

[3] Liu X, Duh K, Gao J. Stochastic answer networks for natural language inference [EB/OL]. [2021-06-25]. https://arxiv.org/pdf/1804.07888v2.pdf.

[4] Kim S, Kang I, Kwak N. Semantic sentence matching with densely-connected recurrent and co-attentive information [C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S. l.: s. n.], 2019.

[5] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Comput, 1997, 9: 1735.

[6] Collobert R, Weston J, Bottou L, *et al*. Natural language processing (almost) from scratch[J]. J Mach Learn Res, 2011, 12: 2493.

[7] Lin Z, Feng M, Santos C N, *et al*. A structured self-attentive sentence embedding [EB/OL]. [2021-06-25]. https://arxiv.org/pdf/1703.03130.pdf.

[8] Gong Y, Luo H, Zhang J. Natural language inference over interaction space[C]//Proceeding of the 6th International Conference on Learning Representations. [S. l.: s. n.], 2018.

[9] Wang Z, Hamza W, Florian R. Bilateral multi-perspective matching for natural language sentences [C]//Proceedings of the Twenty-Sixth International

Joint Conference on Artificial Intelligence. [S. l. : s. n.], 2017.

[10] Tan C, Wei F, Wang W, *et al*. Multiway attention networks for modeling sentence pairs [C]//Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. [S. l. : s. n.], 2018.

[11] Xiong X, Li Y, Zhang R, *et al*. DGI: Recognition of textual entailment via dynamic gate matching [J]. Knowl-Based Syst, 2020, 194: 105544.

[12] Tay Y, Tuan L A, Hui S C . Co-Stack residual affinity networks with multi-level attention refinement for matching text sequences [EB/OL]. [2021-06-25]. https://arxiv. org/pdf/1810. 02938. pdf.

[13] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [EB/OL]. [2021-06-25]. https://arxiv. org/pdf/1409. 0473. pdf.

[14] Wang W, Yan M, Wu C. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering [EB/OL]. [2021-06-25]. https://arxiv. org/pdf/1811. 11934. pdf.

[15] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need [C]//Advances in neural information processing systems. [S. l. : s. n.], 2017.

[16] Zhang Z, Zeng Y, Pang Y. A chinese textual entailment recognition method incorporating semantic role and self-attention [J]. Acta Electonica Sin, 2020, 48: 2162.

[17] Xu J, Cai Y, Wu X, *et al*. Incorporating context-relevant concepts into convolutional neural networks for short text classification [J]. Neurocomputing, 2020, 386: 42.

[18] Bowman S R, Angeli G, Potts C, *et al*. A large annotated corpus for learning natural language inference [C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal. [S. l. : s. n.], 2015.

[19] Khot T, Sabharwal A, Clark P. SCITAIL: A textual entailment dataset from science question answering [C]//Thirty-Second AAAI Conference on Artificial Intelligence. [S. l. : s. n.], 2018.

[20] Williams A, Nangia N, Bowman S R. A broad-coverage challenge corpus for sentence understanding through inference [EB/OL]. [2021-06-25]. https://arxiv. org/pdf/1704. 05426. pdf.

[21] Kingma D P, Ba J. Adam: a method for stochastic optimization [EB/OL]. [2021-06-25]. https://arxiv. org/pdf/1412. 6980v8. pdf.

[22] Srivastava N, Hinton G, Krizhevsky A, *et al*. Dropout: a simple way to prevent neural networks from overfitting [J]. J Mach Learn Res, 2014, 15: 1929.

[23] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: ACL, 2014.

[24] Parikh A, Täckström O, Das D, *et al*. A Decomposable attention model for natural language inference [C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: ACL, 2016.

[25] Yang R, Zhang J, Gao X, *et al*. Simple and effective text matching with richer alignment features [EB/OL]. [2021-06-25]. https://arxiv. org/pdf/1908. 00300v1. pdf.

[26] 高云龙，吴川，朱明. 基于改进卷积神经网络的短文本分类模型[J]. 吉林大学学报：理学版，2020，58：923.

[27] 杨军，王亦民. 基于深度卷积神经网络的三维模型识别[J]. 重庆邮电大学学报：自然科学版，2019，31：253.

引用本文格式：

中　文：王伟，孙成胜，伍少梅，等. 一种轻量级文本蕴含模型[J]. 四川大学学报：自然科学版，2021，58：052001.

英　文：Wang W, Sun C S, Wu S M, *et al*. A lightweight text entailment model[J]. J Sichuan Univ: Nat Sci Ed, 2021, 58: 052001.