基于防退化策略的多通道闭环 BiLSTM 在文本分类中的应用研究 *

孙中宇，龚红仿†，狄俊珂

(长沙理工大学 数学与统计学院, 长沙 410114)

摘 要：双向长短时记忆(BiLSTM)及其变体能够处理可变长度序列，由于文本的复杂语义信息和文本数据嵌入维度的高维性，BiLSTM 表现出低层次网络学习能力较弱，通过叠加网络层学习高层次的特征表示，容易出现网络退化问题。为解决这些问题，提出一种闭环-BiLSTM 模块用于丰富每一层网络结构的隐状态的语义信息表示，同时采用残差连接和增强稀疏表示策略来优化模块，稀疏化隐状态特征向量减缓网络退化问题。最后利用加权融合的多通道词嵌入将语义信息和情感信息在低维张量下实现融合来丰富输入层的文本表示。对情感分类和问题分类的数据集进行了实验验证，实验表明，模型在捕捉文本的情感信息表达具有出色的性能，具有较好的分类精度和鲁棒性。

关键词：文本分类；闭环 BiLSTM；优化策略；多通道词嵌入

中图分类号：TP doi: 10.19734/j.issn.1001-3695.2020.06.0161

Multi-channel closed-loop BiLSTM with anti-degradation strategy for text classification

Sun Zhongyu, Gong Hongfang†, Di Junke

(*School of Mathematics & Statistics*, *Changsha University of Science & Technology*, *Changsha Hunan* 410114, *China*)

Abstract: Bidirectional Long short-term memory(BiLSTM) and its variants can handle variable length sequences. Due to the complex semantic information of text and the high dimension of embedded text data, BiLSTM shows a weak ability of low-level network learning. And the problem of network degradation is easy to occur when learning high-level feature representation using overlay network layer. In order to solve these problems, this paper uses a closed-loop BiLSTM module to enrich the semantic representation of hidden state in each layer of network structure. Meanwhile, it uses residual connection and enhanced sparse representation strategy to optimize the module, and sparse feature vectors of hidden state to alleviate network degradation. Finally, in this paper, to enrich the text representation of the input layer, weighted fusion of multi-channel word embedding is adopted and can be able to realize the fusion of semantic information and emotional information under the low-dimensional tensor. The datasets of emotion classifications and problem classification are verified by experiments. The experiments show that the model in this paper has excellent performance, good classification accuracy and robustness in capturing the expression of emotion information in text.

Key words: text classification; closed-loop BiLSTM; optimizing strategy; multi-channel word embedding

0 引言

随着社交网络的兴起，网络评论成为人们表达情感、分享观点和寻求问题解答的主要方式之一。研究用户在互联网上的情感极性以及评论信息的相关目标实体就具有一定的现实意义[1]。文本分类技术提供了一种自动处理文本信息的方法，现主要用于用户推荐系统、过滤系统等需要识别用户喜好的应用中[2,3]。问题分类和情感极性分类是文本分类的主要内容。前者主要通过提取文本中的语言信息识别文本极性，后者侧重于通过识别文本中的关键字符信息实现目标实体信息提取[4]。传统的文本分类技术只关注了文本中某个或者几个目标词来实现分类问题，这样存在的缺陷是，对有上下文语义联系的短语将会将会出现判断错误。随着深度学习知识在自然语言处理(NLP)、计算机视觉(CV)等领域的广泛应用，模型方法更加注重词语之间的语义联系[5]。

LSTM 是 RNN 增加一种门控单元产生的变体，旨在解决长序列下 RNN 易出现梯度消失和梯度爆炸的问题，是一种顺序序列建模模型[8]。双向 LSTM(BiLLSTM)是 LSTM 网络的进一步扩展，它通过将前向隐状态和后向隐状态结合向下一个网络层中传递，增强文本上下文之间的联系，更好解决文本分类问题[6]。目前，为了学习文本更深层的语义信息表达，普遍采用堆叠更深的神经网络层，从低层次的 n-gram 特征学习到高层次的特征表示，并取得了优异的效果[7]。这样存在一个缺陷是，随着网络层的堆叠，模型精度趋于饱和，网络难以优化。一种残差连接用于堆叠网络层，通过在一层或者几层网络层后施加残差连接，减轻层梯度和传播误差[8]，有效解决了这个问题。由于前向 LSTM 和后向 LSTM 初始隐状态的传递，存在时序信息不足，遗忘门和更新门的算存在偏差；层的堆叠容易使得网络退化，模型训练误差难以收敛。针对这一问题，本文提出一种闭环-BiLSTM 模块(Closed-loop BiLSTM, C-BiLSTM)来丰富前向隐状态和后向隐状态的语义信息表示，减少记忆细胞计算偏差，学习每层网络中更深的情感信息表达。同时避免退化问题，也引入了残差连接策略。

此外，上述模型中主要存在以下两个问题。一方面，C-BiLSTM 模块虽然丰富了每一层网络的隐状态信息表示，但是随着闭环次数的增加，隐状态向量变得不再稀疏，增加了模型过拟合的风险。为此，引入了一种增强稀疏表示策略[9]，通过字典细化选择算法选择具有更加稀疏性的子字典，相比较传统的稀疏算法，该算法具有保留文本原本的语义信息，降低矩阵维度，减少了计算复杂度的优势。另一方面，上述

收稿日期：2020-06-23；**修回日期：**2020-09-01 **基金项目：**国家自然科学基金资助项目(61972055)；湖南省教育厅重点资助项目(18A145)

作者简介：孙中宇(1995-)，男，安徽六安人，硕士，主要研究方向为自然语言处理；龚红仿(1968-)，男(通信作者)，湖北天门人，副教授，硕导，博士，主要研究方向为嵌入式计算系统、信息物理系统、排队论等(gonghf@csust.edu.cn)；狄俊珂(1992-)，女，河南洛阳人，硕士，主要研究方向为最优化理论、排队论.

模型没有为嵌入矩阵提供足够丰富的语义情感信息，这也是目前大部分模型所没有考虑到的问题。大多模型的词嵌入只是单一的考虑了文本编码的语义嵌入，而没有将情感嵌入融合进去。文献[10]明确指出了这个问题，将语义、情感、词语三个多通道嵌入到一个嵌入张量中，构建更加丰富的嵌入表示。但是对于上述嵌入矩阵，会增加嵌入矩阵维度，不利于模型训练。本文采用一种加权融合的多通道词嵌入，保证多通道的词嵌入维度媲美于单个词嵌入维度时，又融合了每种词嵌入的语义情感信息表示。

基于上述已有的工作总结，本文提出一种带有增强稀疏表示和残差连接的多通道闭环 BiLSTM 模型(E⁺RMC-BiLSTM)用于文本分类。该模型由多通道嵌入层，闭环 BiLSTM 模块(C-BiLSTM)，增强稀疏表示和残差连接策略组成。本文主要贡献有如下 3 点：

a)采用加权融合的多通道词嵌入，降低多通道的嵌入矩阵维度同时，融合了语义情感信息表示，避免了单一词嵌入导致情感信息丢失。

b)C-BiLSTM 模块丰富了每层网络的前向隐状态和后向隐状态的语义信息表示，在每层网络中尽可能多的学习到高层特征信息表示，减少门控单元的计算偏差。

c)增强稀疏表示策略稀疏化了隐状态向量，降低了模型过拟合训练集的风险；残差连接策略减轻了多层网络叠加时，出现的训练误差难以收敛的问题。

1　相关工作

随着深度学习领域的快速发展，文本分类问题作为 NLP 任务的一个分支，已经被广泛研究。LSTM 模型因其在处理长序列数据的优异性能，而被广泛应用。目前，各种 LSTM 的变体已经被提出，满足不同的任务需求，实现更加智能的情感识别[11]。

LSTM 的变体大多旨在增强目标实体与上下文之间的联系来实现更加精准文本分类，结合其他网络模型和优化每一层状态输出是一个重要的研究方向。申静波等人[12]在 LSTM 的隐含层和输出层之间对重要句子引入注意力机制有效提升文本的极性分类。Shuang 等人[13]针对情感分类的三个关键因素，设计了两种情感分析模块，增强语境词的位置，距离等之间的联系，提高分类精度。Abdi 等人[14]提出一种 LSTM 变体 RNSA，利用情感知识，情感转移和规则克服了传统方法中词的顺序和信息消息缺陷的问题。

近年来，残差连接在解决层堆叠出现的退化问题具有优异的性能表现。Conneau 等人[15]堆叠 29 个卷积层，构建一个深层的卷积网络用于文本分类，实验证实了堆叠神经网络在 NLP 任务中具有一定的优势。Wang 等人[8]针对八层 BiLSTM 退化问题引入残差连接策略和平均池化用于情感强度预测，实验发现退化八层模型性能发生了逆转。Cao 等人[16]开发了一种密集的递归 CNN 模块，并采用基于残差的短连接体系结构，以分层集成多级特征图，实验发现显著减少的模型参数可实现更精准的效果。目前基于稀疏表示(SR)用于文本分类的文献研究较少，而且一些早期的稀疏方法文献并不适用文本分类。Sainath 等人[17]提出一种扩展稀疏表示方法，发现基于最大 l_2 支持的文本分类方法优于其他 SR 方法，并在一定范围内，分类性能优于 Nave Bayes 分类器。Sharma 等人[18]提出一种稀疏表示分类器和支持向量机的组合分类器，并使用基于文本的词频表示进行文本分类。Yao 等人[19]采用随机投影降维，并利用稀疏表示获取语义相关性的稀疏解。Unnikrishnan 等人[9]提出一种字典细化过程保证特征稀疏的同时，特征之间的冗余度与噪声更低。

最近，在一些情感语义分析的研究中，词嵌入方式被证明是可以有效提高文本的分类性能。Stein 等人[20]将经典的 Glove，Word2Vec 和 fastText 三种词嵌入用于特定的层次化文本分类问题中，发现 Word2Vec 嵌入取得了较好的效果。Guo 等人[21]针对不同类别标签中词的重要性赋予不同权值应用于词嵌入。Wu 等人[22]基于连接的词嵌入(CBWE)学习对显示话语数据进行连接分类，学习的 CBWE 能够捕捉词与词的话语关系，作为预先训练好的词嵌入，用于隐性话语识别。Zhang 等人[23]将语义嵌入，情感嵌入和词典嵌入用于文本编码，并结合注意向量，LSTM 注意和注意池三种注意方法与 CNN 集成在一起用于文本情感分析。文献[10]在文献[23]基础上进行了扩展，将三种嵌入方式合并在一个嵌入张量中丰富嵌入层的语义表示。

在上述的研究工作中，本文从 LSTM 变体、残差连接、稀疏表示、词嵌入四个方面阐述了目前的研究现状，内容都有所交叉。这些模型方法都是本文 E⁺RMC-BiLSTM 模型的基础。其中后文提到的 MC-BiLSTM 模型是丢失增强稀疏表示和残差连接两种策略的模型，定义为多通道闭环-BiLSTM 模型；C-BiLSTM 是丢失上述两种策略和多通道词嵌入的模型，定义为闭环-BiLSTM 模型，这些模型都是本文 E⁺RMC-BiLSTM 模型的子模块模型。

2　系统模型

本节将介绍带有增强稀疏表示和残差连接的多通道闭环 BiLSTM 的系统模型(图 1)。首先，对文本数据采用多通道的词嵌入处理，丰富文本的语义情感信息表示；接着，闭环 BiLSTM 模型用于提高对文本情感信息的识别能力，同时在闭环 LSTM 的每个隐状态部分施加残差连接策略与增强稀疏表示策略，提高隐状态的信息表示能力，最终在输出层通过线性解码器预测指定目标实体的情感类别概率，确定最终情感类别。

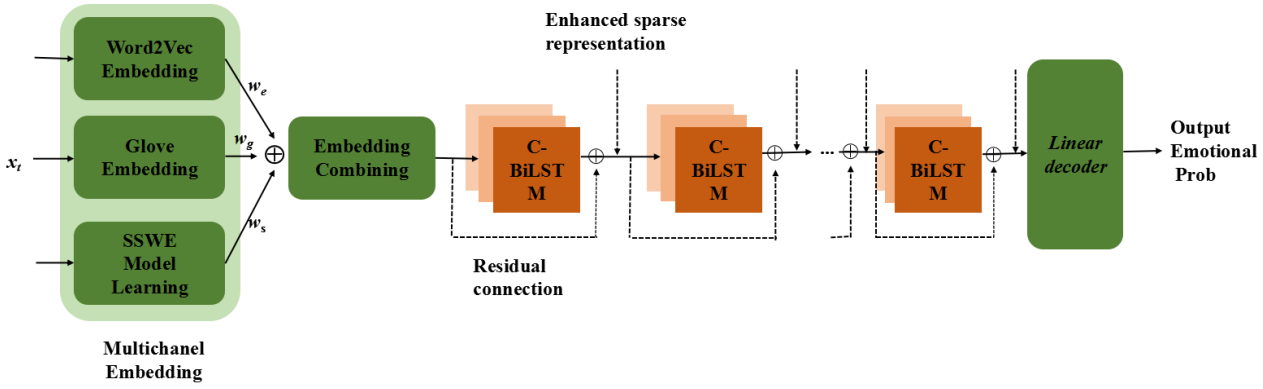

图 1　系统模型

Fig. 1　The model of system

2.1　多通道嵌入

本节将介绍一种加权融合的多通道嵌入，在降低嵌入矩阵维度的同时，有效增强词向量的语义信息与情感表达。一个训练文本每次截取 N 个单词，vr_n 是训练文本中第 n 个单词向量，存在嵌入矩阵 W_e,W_g,W_s 将单词嵌入为一种词向量表示 L_{en}、L_{gn}、L_{sn}，式(1)表示如下：

$$L_{en}=W_e vr_n \quad L_{gn}=W_g vr_n \quad L_{sn}=W_s vr_n \tag{1}$$

其中 $n\in[1,N]$,We,W_g,W_s 分别是 *Word2vec*、*Glove* 词嵌入和 *SSWE* 模型学习到的语义向量作为感情表示。

为了能够降低嵌入矩阵维度，同时保持文本的语义信息与情感表示，采用加权融合方式，式(2)表示如下：

$$L_{cn}=w_e L_{en}\oplus w_g L_{gn}\oplus w_s L_{sn} \tag{2}$$

其中 $w_i(i=e,g,s)$表示每个分布式词嵌入所占有的权重大小，$\oplus$ 表示词向量对应元素相加。

2.2　C-BiLSTM

本节介绍了一种闭环-双向 LSTM 模型提升一个 BiLSTM 层中的初始隐状态的文本信息量，同时记忆细胞多次判断文本信息重要性，减少遗忘门与更新门的计算误差。

图 2 表示了闭环-BiLSTM 的算法流程图。首先将加权融合的多通道嵌入词向量 Lc_n 传入前向 LSTM 中，并将产生的每个时间步的隐状态向下一个神经单元中传递，前向 LSTM 中最后一个神经单元产生的隐状态 ${}^k\vec{h}_N^l$ 传递给后向 LSTM 的初始神经单元中，替换初始隐状态 ${}^k\overleftarrow{h}_0^l$，接着后向 LSTM 向下依次传递更新迭代的隐状态特征；同理，后向 LSTM 中最后一个神经单元产生的隐状态 ${}^k\overleftarrow{h}_N^l$ 将会替换前向 LSTM 的初始隐状态 ${}^k\vec{h}_0^l$，并进行下一个时间步的迭代更新，从而形成闭环-BiLSTM。最后通过控制闭环-BiLSTM 的循环次数，将前向 LSTM 与后向 LSTM 的最终隐状态特征传递给下一个网络层预测文本情感极性(也可以是多层叠加的闭环-BiLSTM 层)。闭环-BiLSTM(C-BiLSTM)计算式(3)(4)表示如下：

$$\begin{aligned}
&\text{if}\quad l=1:\\
&{}^k\vec{h}_t^l=\vec{LSTM}(Lc_t,{}^k\vec{h}_{t-1}^l),t\in[1,N]\\
&{}^k\overleftarrow{h}_0^l={}^k\vec{h}_N^l,if\quad t=N\\
&\text{if}\quad l>1:\\
&{}^k\vec{h}_t^l=\vec{LSTM}({}^k\vec{h}_t^{l-1},{}^k\vec{h}_{t-1}^l),t\in[1,N]\\
&{}^k\overleftarrow{h}_0^l={}^k\vec{h}_N^l,if\quad t=N
\end{aligned} \tag{3}$$

$$\begin{aligned}
&\text{if}\quad l=1:\\
&{}^k\overleftarrow{h}_t^l=\overleftarrow{LSTM}(Lc_t,{}^k\overleftarrow{h}_{t+1}^l),t\in[N,1]\\
&{}^k\vec{h}_0^l={}^k\overleftarrow{h}_1^l,if\quad t=1\\
&\text{if}\quad l>1:\\
&{}^k\overleftarrow{h}_t^l=\overleftarrow{LSTM}({}^k\overleftarrow{h}_t^{l-1},{}^k\overleftarrow{h}_{t-1}^l),t\in[N,1]\\
&{}^k\vec{h}_0^l={}^k\overleftarrow{h}_1^l,if\quad t=1
\end{aligned} \tag{4}$$

其中 ${}^k\vec{h}_t^l$、${}^k\overleftarrow{h}_t^l$ 分别是前向 LSTM 和反向 LSTM 的第 k 次循环的第 l 层的第 t 个时间步的隐状态特征，$\vec{LSTM}$、$\overleftarrow{LSTM}$ 是 LSTM 的内部单元的计算方式，返回当前时间步的隐状态。

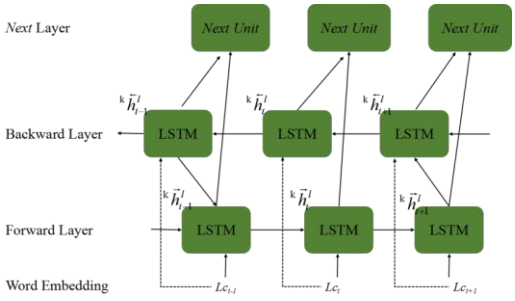

图 2　C-BiLSTM 模块细节

Fig. 2　Details of C-bilstm module

2.3　残差连接与增强稀疏表示策略

本节针对随着闭环-BiLSTM 循环次数的增加，每层神经单元产生的隐状态特征包含的语义信息越来越丰富，隐状态向量将不再具有稀疏性这一问题，引入残差连接与增强稀疏表示两种策略，在进一步提高隐状态的语义表达能力的同时，稀疏化文本词向量的特征表示。

图 3 表示了残差连接与增强稀疏表示策略的流程图。将前向 LSTM 和后向 LSTM 产生的当前时间步的隐状态分别与 m 层输入层向量 Lc_t 对应元素相加，从而通过残差连接策略学习到的新的隐状态由式(5)(6)表示：

$${}^k\vec{h}_t^l=\vec{LSTM}(Lc_t,{}^k\vec{h}_{t-1}^l)\oplus Lc_t^{l-m},t\in[1,N] \tag{5}$$

$${}^k\overleftarrow{h}_t^l=\overleftarrow{LSTM}(Lc_t,{}^k\overleftarrow{h}_{t-1}^l)\oplus Lc_t^{l-m},t\in[N,1] \tag{6}$$

其中 m 是残差连接中 C-BiLSTM 中间层的数量。

通过将残差连接处理后的前向隐状态和后向隐状态进行拼接形成更深维度的状态特征 $H_t=\left[{}^k\vec{h}_t^l,{}^k\overleftarrow{h}_t^l\right]$，将其视为一个字典特征。接着，采用基于聚类算法的字典细化过程(CDPR)生成新的字典特征 $\hat{H}_t$，以此保证字典特征之间不具有相似性。通过皮尔逊相关系数 r 作为度量指标，选择 k 个与新的字典特征 $\hat{H}_t$ 最为相似的子字典，这个过程即增强稀疏表示[9]，对于筛选掉的特征向量全都用 0 进行填充。稀疏化后的隐状态向量表示如下：

$$\hat{H}_t=CDRP(H_t)=CDRP\left[{}^k\vec{h}_t^l,{}^k\overleftarrow{h}_t^l\right] \tag{7}$$

$$\hat{H}_t^E=E^+\text{-}SR(\hat{H}_t) \tag{8}$$

其中 CDRP、E⁺-SR 分别是字典细化过程和增强稀疏表示的算法，分别返回新的字典特征和进一步稀疏的字典特征。值得注意的是，为了能够将稀疏后的状态特征向下一个映射层中传递，在文献[9]的算法 4 中没有使用 *argmin* 函数，而是直接返回整个的字典特征作为稀疏化的隐状态特征，即 E⁺SRC 算法退化为 E⁺SR。

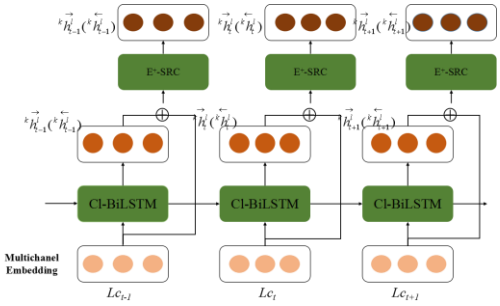

图 3　残差连接与增强稀疏表示策略

Fig. 3　Residual connections and enhanced sparseness represent policies

最后在多层 C-BiLSTM 层之后添加 *dropout* 层和线性解码器，输出文本的情感类别概率，确定文本的感情极性。

为了能够有效评价分类问题的模型性能，本文采用交叉熵损失函数作为评价函数，并施加 L_2 正则化，其值越小，表明越拟合真实数据部分，在反向误差传播中采用 RMSprop 优化器优化模型参数。交叉熵的损失函数公式如下：

$$Loss=-\frac{1}{M}\sum_M[y_{true}\ln y_{pred}+(1-y_{true})\ln(1-y_{pred})]+\lambda\|w\|_2^2 \tag{9}$$

其中 M 是训练样本的大小，y_{true} 表示真实样本的情感类别，y_{pred} 是预测的感情类别，λ 是 L_2 正则化系数。

3　实验

3.1　数据集

为了验证本文模型的有效性，本文选取了四组公共数据集，分别是三个情感分类数据集 MR，STT-2 和 IMDB 和分

类问题数据集 TREC。数据集汇总信息如下。

a) MR：电影感情极性评论数据集，一共有 5331 个正面评论和 5331 个负面评论，需要能够正确划分正面与负面评论。

b) SST-2：对 SST-1 数据集进行标注的二分类数据集，删除 SST-1 数据集中的中性标注，并将积极和非常积极统一标注为积极标签，将消极和非常消极标注为消极标签。

c) IMDB：电影情感评论数据集，训练集和测试集均有 25000 条评论数据，要求对测试集正确判断评论信息是积极或消极。

d) TREC：电影长序列情感评论数据集，实现积极或消极分类任务。

表 1 是文本数据集的汇总信息。其中 *Dict* 表示每个数据集的文本所包含的词汇量大小，

L 是每条评论数据中平均单词个数，*Class* 表示分类任务，*5_KFlods* 表示没有测试集，本文对训练集作了五折交叉验证处理。

表 1　文本数据集汇总信息

Tab. 1　Summary information of text datasets

Dataset	Train	Test	Dict	L	Class	Class Type
MR	10662	5_KFlods	18765	20	2	sentiment
SST-2	7801	1812	16185	19	2	sentiment
TREC	5452	500	9592	10	6	question
IMDB	25000	25000	392000	231	2	sentiment

3.2　实现细节

本文与现有的基线方法进行了比较，这些基线方法在当时都具有最优的分类性能。本文从机器学习、卷积神经网络和递归神经网络及其变体、词嵌入方式四个角度描述每个模型算法的有关细节。

a) 机器学习：支持向量机(SVM)；多项式朴素贝叶斯(MNB)；朴素贝叶斯 log-count ratios 作为特征值的 SVM 变体(NBSVM)；无监督的段落向量算法学习可变文本长度的特征表示(Paragraph-Vec)；

b) CNN 及其变体：提出动态 k-max 池化层的动态 CNN 用于建模句子(DCNN)[7];*Word2Vec* 预训练词嵌入(CNN-static)并使用了微调优化策略的 CNN 模型(CNN-non-static)以及两套预训练词嵌入的一维 CNN(CNN-multichannel)。

c) RNN,LSTM 及其变体：CNN 与 RNN 结合的文本分类模型(C-LSTM);提出了一种多任务共同学习的 LSTM 框架(Multi-task LSTM)[8]；将 LSTM 改进为树状的网络拓扑结构(Tree-LSTM)[11]；带有短语因子机制，提取更丰富的文本信息的 LSTM 模型(P-LSTM)；利用解析树上每个节点分支使用向量和矩阵的 RNN(MV-RNN)；基于张量特征函数的 RNN(RNTN)；多层递归叠加的 RNN 模型(DRNN)；提出带有注意力机制和一维 CNN 层的 BiLSTM 模型[6]。

d) 词嵌入方式：利用 Wikipedia 预训练的词嵌入方式(RAE)。

e) E⁺RMC-BiLSTM：本文提出的 E⁺RMC-BiLSTM 模型。设计了闭环双向 LSTM 闭环次数为 1(设置为 Type1)和闭环次数为 2(设置为 Type2)两种模块。本文没有尝试闭环次数为 3 及以上的情况，因为这对于计算复杂度和时间开销而言都是不可取的。

3.3　实验参数设置

针对本文提出的 E⁺RMC-BiLSTM 模型，实验参数中每个词嵌入方法的嵌入大小均为 500，加权融合嵌入分别为 0.2,0.2,0.6 和 0.3,0.3,0.4 两种权重组合；BiLSTM 内存维度设为 300；dropout 正则化率为 0.6；梯度下降采用 RMSProp 优化器；学习速率 0.01；随机初始化参数，满足正态分布于 (−0.01,0.01) 之间。采用验证集的精确度作为模型的衡量指标。对于其简化模块的几种模型，具有该模块的模型参数都是一致的。

3.4　评价 C-BiLSTM 模块性能

为了验证闭环模块次数对系统模型的影响。在保证具有相同加权融合多通道词嵌入、残差连接和增强稀疏表示策略下，分别比较了四层 C-BiLSTM 和八层 C-BiLSTM 的闭环次数为 1 或者 2，以及单一的八层 BiLSTM 模块的性能差异。

在图 4 中可以发现，四层 C-BiLSTM 的闭环次数为 1 时，性能要优于没有闭环模块的八层的 BiLSTM。这意味着丰富一个 C-BiLSTM 层中隐状态的情感信息表达比堆叠 BiLSTM 层获取深层的文本语义情感信息而言，具有更好性能。同时，层的堆叠使得网络容易出现梯度消失或者梯度爆炸的问题，也更容易使得网络的训练误差难以收敛。另一方面，闭环次数的增加，模型的整体分类性能都有所增加，也就是说 C-BiLSTM 模块有效的增强了记忆细胞的判断能力，减少了遗忘门与更新门的计算误差，提升了系统模型的分类精度。

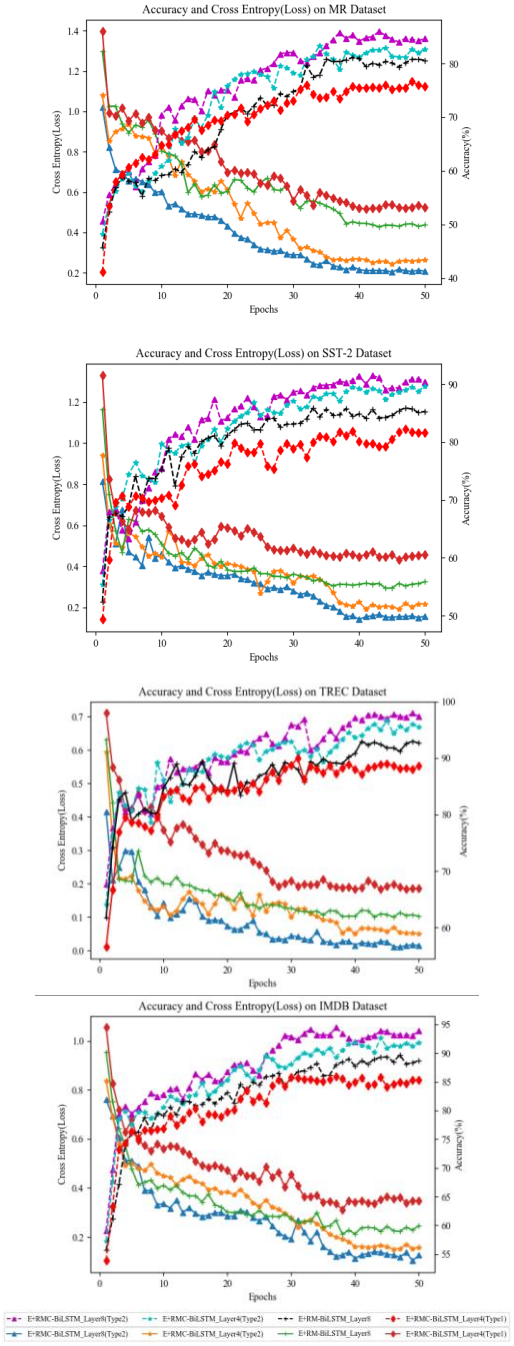

图 4　模型在不同闭环次数下的交叉熵损失值与精度变化

Fig. 4　The changes of model cross-entropy loss value and accuracy in different closed loops

3.5　评价两种优化策略

在本节中本文分析在有无残差连接和增强稀疏表示两种策略下，系统模型的性能变化。图 5 展现的是单一的四层和八层 MC-BiLSTM 以及带有两种策略的四层和八层 MC-BiLSTM，其他参数不变(闭环类型均是 Type2)，模型的交叉熵损失值和精度变化。

图 5 中可以发现，单一八层 MC-BiLSTM 由于层的堆叠使得网络难以训练，泛化能力甚至低于单一的四层 MC-BiLSTM。随着残差连接和增强稀疏表示策略层的添加，向每一层神经单元传递的隐状态特征变得更加稀疏，有效解决了训练误差收敛速度和梯度消失两个问题，模型性能发生了逆转。添加两种策略的模型性能都要优于未施加策略的模型，在 MR、IMDB、SST-2、TREC 四个数据集上，添加策略后的八层 MC-BiLSTM 相对与单一的八层 MC-BiLSTM 相对性能分别提升了 2.90%、2.05%、2.44%、3.56%。

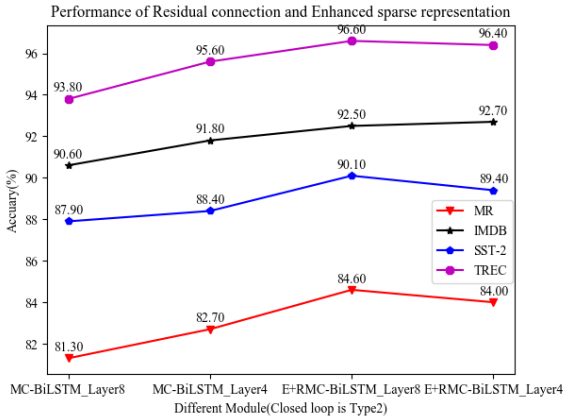

图 5　两种优化策略的性能评价

Fig. 5　Performance reviews of two optimize strategies

3.6　评价多通道词嵌入模块

为了考察不同词嵌入方式对系统模型的性能影响。在相同的 E⁺RMC-BiLSTM 系统模型下(Type2)，比较了 *Word2vec*(预训练法)、*GloVe*、*SSWE* 以及不同加权融合的词嵌入方法。采用 F_1-*score* 来衡量二分类数据集的性能，它是由 Recall 和 Precision 构成，这三个指标越高，代表模型的分类性能和鲁棒性越好[6]，F_1-*score* 的计算公式由式(10)表示。

$$F_1\text{-}score = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{10}$$

表 2 中可以发现，带有不同加权方式融合的多通道词嵌入在三组二分类数据集上具有最优的 F1-score。其他三种单一的词嵌入方在不同的数据集上有不同的性能表现，其中预训练的 *Word2Vec* 和 *SSEW* 模型的两种嵌入方式表现较为出色。甚至在 IMDB 数据集中，预训练的 *Word2Vec* 精度高于多通道的词嵌入(we=0.3,w_g=0.3,w_s=0.4)。但是采用不同的加权权重实现嵌入特征融合，有效的提升了模型的整体分类性能。

3.7　全面比较

在本节中实验了情感分类(长评论 IMDB、短评论 MR,SST-2)以及问题分类(TREC)四种数据集上的模型精度变化。表 3 中给出了 17 种机器学习和深度学习算法，具有当时最优的分类性能。从下表中可以发现，四层、八层 E⁺RMC-BiLSTM 在情感数据集上(MR、SST-2、IMDB)具有最优的分类性能。相比较基线方法中最优的分类结果，相对性能分别提升了 1.65%、2.00%、0.97%。在问题分类数据集 TREC 中，本文的 E⁺RMC-BiLSTM 模型略差于 Liu 等人[6]提出的 AC-BiLSTM 模型，但是优于大部分基线模型。在一定程度上说明了，本文的系统模型在捕捉文本的情感信息表达具有较强的能力，同时在问题理解分类任务中也具有一定优势。

表 2　不同的词嵌入性能比较

Tab. 2　Performance comparison of different word embedding

DataSet	Embedding Methods	Recall	Precision	F_1-score
MR	*Word2Vec*(*pre-trained*)	82.15	80.74	81.44
	Glove	77.84	79.30	78.56
	SSEW	81.22	83.46	82.32
	Multi-channel word embedding (we=0.2,w_g=0.2,w_s=0.6)	83.32	84.14	83.73
	Multi-channel word embedding (we=0.3,w_g=0.3,w_s=0.4)	85.17	84.51	**84.84**
SST-2	*Word2Vec*(*pre-trained*)	81.95	83.71	82.82
	Glove	80.11	82.82	81.44
	SSEW	83.46	85.14	84.29
	Multi-channel word embedding (we=0.2,w_g=0.2,w_s=0.6)	89.45	91.67	**90.55**
	Multi-channel word embedding (we=0.3,w_g=0.3,w_s=0.4)	86.34	88.63	87.47
IMDB	*Word2Vec*(*pre-trained*)	90.33	91.74	91.03
	Glove	85.67	83.72	84.68
	SSEW	88.96	90.76	89.85
	Multi-channel word embedding (we=0.2,w_g=0.2,w_s=0.6)	91.52	92.50	**92.01**
	Multi-channel word embedding (we=0.3,w_g=0.3,w_s=0.4)	92.83	90.96	91.89

表 3　比较不同的方法在情感分类和问题分类数据集上的精度

Tab. 3　Comparision of different methods on emotional classification and question classification datasets

Methods	MR	SST-2	TREC	IMDB
SVM	-	79.4	95	89.2
MNB	79.0	-	-	86.6
NBSVM	79.4	-	-	91.2
RAE	77.7	82.4	-	-
MV-RNN	79.0	82.9	-	-
RNTN	-	85.4	-	
Paragraph-Vec	-	87.8	91.8	-
DCNN	-	86.8	93	-
CNN-static	81.0	86.8	-	-
CNN-non-static	81.5	87.2	93.6	-
CNN-multichannel	81.1	88.1	92.2	-
DRNN	-	86.6	-	-
C-LSTM	-	-	94.6	-
Multi-task LSTM	-	87.9	-	-
Tree-LSTM	-	86.9	-	-
P-LSTM	-	-	-	91.5
AC-BiLSTM	83.2	88.3	**97.0**	91.8
E⁺RMC-BiLSTM_Layer8(Type1)	83.4	88.9	95.3	92.2
E⁺RMC-BiLSTM_Layer4(Type1)	83.1	88.7	94.7	91.1
E⁺RMC-BiLSTM_Layer8(**Type2**)	**84.6**	**90.1**	96.6	92.5
E⁺RMC-BiLSTM_Layer4(**Type2**)	84.0	89.4	96.4	**92.7**

4　结束语

本文提出了一种带有增强稀疏表示和残差连接的多通道闭环 BiLSTM 来实现文本分类。C-BiLSTM 模块丰富了每层 LSTM 网络中隐状态的语义信息表示，残差连接和增强稀疏表示分别起到了防止多层堆叠的 C-BiLSTM 模块退化问题和稀疏化隐状态的特征向量防止过拟合训练集的作用。加权融合的多通道词嵌入则丰富了嵌入层的语义信息和情感信息。本文在三个情感数据集和一个问题分类数据集上，与现有的

基线方法相比，表明本文模型在捕捉文本的情感信息表达具有出色的性能，并取得了较好的分类精度和鲁棒性。

未来研究工作主要是 BiLSTM 结构的优化和词嵌入方式的研究，主要包括以下几个部分：a)研究其他更加简洁有效的策略，优化 BiLSTM 的状态输出；b)尝试更加丰富有效的词嵌入以及词嵌入融合；c)设计新的网络结构结合 BiLSTM 用于更加细腻的情感分析。

参考文献：

[1] Xu Jie, Huang Feiran, Zhang Xiaoming, *et al.* Visual-textual sentiment classification with bi-directional multi-level attention networks [J]. Knowledge-Based Systems, 2019, 178: 61-73.

[2] Watanabe A, Sasano R, Takamura H, *et al.* Generating Personalized Snippets for Web Page Recommender Systems [C]// Proc of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) , Warsaw, 2014: 218-225.

[3] Roy P K, Singh J P, Banerjee S. Deep learning to filter SMS Spam [J]. Future Generation Computer Systems, 2020, 102: 524-533.

[4] Liu Bing. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies [M]. San Rafael: Morgan and Claypool Publishers, 2012: 50-160.

[5] 张洋, 胡燕. 基于多通道深度学习网络的混合语言短文本情感分类方法 [J]. 计算机应用研究, 2021, 38 (1): 220-227. (Zhang Yang, Hu Yan. Emotion classification method of mixed language short text based on multi-channel deep learning Network [J]. Application Research Of Computers, 2021, 38 (1): 220-227.)

[6] Liu Gang, Guo Jiabao. Bidirectional LSTM with attention mechanism and convolutional layer for text classification [J]. Neurocomputing, 2019, 337: 325-338.

[7] Farabet C, Couprie C, Najman L, *et al.* Learning Hierarchical Features for Scene Labeling [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35 (8): 1915-1929.

[8] Wang Jin, Peng Bo, Zhang Xuejie, Using a stacked residual LSTM model for sentiment intensity prediction [J]. Neurocomputing, 2018, 322: 93-101.

[9] Unnikrishnan P, Govindan V K, Kumar S D M. Enhanced sparse representation classifier for text classification [J]. Expert Systems with Application, 2019, 129: 260-272.

[10] Gan Chenquan, Wang Lu, Zhang Zufan, *et al.* Sparse attention based separable dilated convolutional neural networkfor targeted sentiment analysis [J]. Konwledge-Based Systems, 2019, 188: 104827.

[11] Tai K S, Socher R, Manning C D. Improved semantic representations from tree-structured long short-term memory networks [C]// Proc of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing ACL-IJCNLP 2015, Association for Computational Linguistics (ACL) , Beijing, China, 2015: 1556-1566.

[12] 申静波, 李井辉, 孙丽娜. 注意力机制在评论文本情感分析中的应用研究 [J]. 计算机技术与发展, 2020, 30 (07): 169-173. (Shen Jingbo, Li Jinghui, Sun Lina. Application of attention mechanism in emotion analysis of critical texts [J]. Computer Technology and Development, 2020, 30 (07): 169-173.)

[13] Shuanga K, Rena X, Yanga Q, *et al.* AELA-DLSTMs: Attention-Enabled and Location-Aware Double LSTMs for aspect-level sentiment classification [J]. Neurocomputing, 2019, 334: 25-34.

[14] Abdi A, Shamsuddin S M, Hasan S, *et al.* Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion [J]. Information Processing & Management, 2019, 56 (4): 1245-1259.

[15] Conneau A, Schwenk H, Cun Y L, *et al.* Very deep convolutional networks for text classification [C]// Proc of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL) , 2017: 1107-1116.

[16] Cao Yanpeng, Fu Guizhong, Yang Jiangxin, *et al.* Accurate salient object detection via dense recurrent connections and residual-based hierarchical feature integration [J]. Signal Processing: Image Communication, 2019, 78: 103-112.

[17] Sainath T N, Maskey S R, Kanevsky D, *et al.* Sparse representations for text categorization [C]// Proc of the 11th Annual Conference of the International Speech Communication Association, International Speech Communication Association (ISCA) , Makuhari, Chiba, Japan, 2010: 26-30.

[18] Sharma N, Sharma A, Thenkanidiyoor V, *et al.* Text classification using combined sparse representation classifiers and support vector machines [C]// Proc of 2016 4th International Symposium on Computational and Business Intelligence (ISCBI) , Computational and Business Intelligence (CBI) , Olten, Switzerland, 2016: 181-185.

[19] Yao L, Sheng Q Z, Wang X, *et al.* Collaborative text categorization via exploiting sparse coefficients, World Wide Web. 2018, 21 (2): 373-394.

[20] Stein R A, Jaques P A, Valiati J F. An analysis of hierarchical text classification using word embeddings [J]. Information Sciences, 2019, 471: 216-232.

[21] Guo Bao, Zhang Chunxia, Liu Junmin, *et al.* Improving text classification with weighted word embeddings via a multi-channel TextCNN model [J]. Neurocomputing, 2019, 363: 366-374.

[22] Wu Changxing, Su Jingsong, Chen Yidong, *et al.* Boosting implicit discourse relation recognition with connective-based word embeddings [J]. Neurocomputing, 2019, 369, 39-49.

[23] Zhang Zufan, Zou Yang, Gan Chenquan. Textual sentiment analysis via three different attention convolutional neural networks and cross-modality compriseent regression [J]. Neurocomputing, 2018, 275: 1407-1415.