

基于注意力 CNLSTM 模型的新闻文本分类

刘 月, 翟东海, 任庆宁

(西南交通大学 信息科学与技术学院, 成都 610097)

摘 要: 结合卷积神经网络(CNN)和嵌套长短期记忆网络(NLSTM)2 种模型, 基于注意力机制提出一个用于文本表示和分类的 CNLSTM 模型。采用 CNN 提取短语序列的特征表示, 利用 NLSTM 学习文本的特征表示, 引入注意力机制突出关键短语以优化特征提取的过程。在 3 个公开新闻数据集中进行性能测试, 结果表明, 该模型分类准确率分别为 96.87%、95.43% 和 97.58%, 其性能比 baseline 方法有显著提高。

关键词: 卷积神经网络; 特征表示; 嵌套长短期记忆网络; 注意力机制; 文本分类

中文引用格式: 刘月, 翟东海, 任庆宁. 基于注意力 CNLSTM 模型的新闻文本分类[J]. 计算机工程, 2019, 45(7): 303-308, 314.

英文引用格式: LIU Yue, ZHAI Donghai, REN Qingning. News text classification based on CNLSTM model with attention mechanism[J]. Computer Engineering, 2019, 45(7): 303-308, 314.

News Text Classification Based on CNLSTM Model with Attention Mechanism

LIU Yue, ZHAI Donghai, REN Qingning

(School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610097, China)

【Abstract】 Combining Convolutional Neural Network (CNN) and Nested Long Short-Term Memory (NLSTM) models, this paper proposes a CNLSTM model for text representation and classification based on the attention mechanism. The model uses CNN to extract feature of phrase sequences, and then uses NLSTM to learn the representation of text features. By introducing attention mechanisms, the key phrases are highlighted to optimize feature extraction. Experiments on three published news data sets demonstrate that the classification accuracy of the model is 96.87%, 95.43%, and 97.58%, respectively, and its performance is significantly improved compared with the baseline methods.

【Key words】 Convolutional Neural Network (CNN); feature representation; Nested Long Short-Term Memory (NLSTM); attention mechanism; text classification

DOI: 10.19678/j.issn.1000-3428.0051312

0 概述

文本分类是自然语言处理的热点和关键技术之一, 在许多实际系统如 Web 内容管理、搜索引擎、邮件过滤等中都扮演着至关重要的角色。而文本分类所要解决的首要问题是捕捉不同文本单元的特征, 如短语、句子和文档等。

由于具有捕捉局部空间或时间结构相关性的能力, 卷积神经网络(Convolutional Neural Network, CNN)自提出以来就受到广泛的关注和应用, 例如在自然语言处理领域。文献[1]提出一种结合情感词典和 CNN 的情感分类方法, 将 CNN 提取的抽象词语序列特征用于情感极性分类; 文献[2]提出一种基于事件卷积特征的新闻文本分类方法, 即通过 CNN 提取文本的事件特征并用于文本分类。然而, CNN 在处理文本

时仍然存在局限性, 它没有考虑距离较远的词语之间的联系, 忽视了语言中依存关系的结构特点。

文献[3]提出利用基于时间序列的循环神经网络(Recurrent Neural Network, RNN)进行文本分类。RNN 是一种具有“记忆”功能的网络模型, 通过链式神经网络架构传播历史信息, 因此能够捕获序列的长期依赖关系。但实际上, 随着 2 个时间步长之间的差距变大, 标准 RNN 也无法学习长期依赖性。文献[4]提出的长短期记忆网络(Long Short-Term Memory, LSTM)从根本上解决了长期依赖问题。当前, LSTM 主要通过堆叠的方式构造多层前馈网络来处理数据, 上一层输出为下一层的输入。而文献[5]提出一种嵌套 LSTM(Nested LSTM, NLSTM), 其含有多层记忆单元, 通过嵌套而不是堆叠的方式增加 LSTM 的深度。NLSTM 可以选择性

基金项目: 国家自然科学基金(61540060)。

作者简介: 刘 月(1993—), 女, 硕士研究生, 主研方向为数据挖掘、自然语言处理; 翟东海, 副教授、博士; 任庆宁, 硕士研究生。

收稿日期: 2018-04-23 **修回日期:** 2018-06-07 **E-mail:** 912895630@qq.com

地访问内部记忆,这使得内部记忆能够在更长的时间尺度上记忆和处理事件,即使这些事件与当前事件无关。因此,与简单的 LSTM 和堆叠式 LSTM 相比较,NLSTM 能够处理更长时间尺度的历史信息。

本文结合 CNN 与 NLSTM 2 种结构的优点,构造一个 CNLSTM 混合体系结构,将基于词语序列的简单 NLSTM 模型扩展到基于短语序列的混合模型以学习文本特征。由于新闻文本通常为长文本,因此文本中可能存在与新闻主题无关的信息。通过在 NLSTM 后引入注意力机制计算注意力概率分布,获得具有短语重要性区分度的文本特征表示。

1 相关工作

1.1 文本分类

文本分类的基本过程为文本预处理、特征选择、分类器训练及结果评估,目的是将文档归类到一组预定义的类中^[6]。传统的文本分类通常采用机器学习算法,如 K 近邻 (K-Nearest Neighbor, KNN)^[7]、朴素贝叶斯^[8]和支持向量机 (Support Vector Machine, SVM)^[9]等,这些方法通常存在维度高、数据稀疏性问题。神经网络的出现为解决数据稀疏性问题提供了新的思路,并提出了许多学习词语特征表示的神经模型,如文献[10]用于文本建模的卷积模型和文献[11]概率语言模型。

中文文本与英文文本分类的一个重要差别在于预处理阶段,中文文本的读取需要分词,而英文文本直接通过单词间的空格区分词语。从简单的查词典方法,到基于统计语言模型的分词方法,中文分词的技术已趋于成熟,很多开源分词器可直接对中文进行分词,如 jieba、中国科学院计算所开发的汉语词法分析系统 ICTCLAS 等。本文使用 jieba 分词器对文本数据进行分词。

1.2 词向量

要将自然语言理解的问题转化为机器学习的问题,第一步将自然语言数学化,转化为计算机能够识别和处理的形式,即将文本的表达映射到 k 维向量空

间^[12],用词语的分布式表示法——词向量表示词语。

本文使用 Skip-gram 模型训练词语的连续词向量表示。Skip-gram 模型是一个带有单层隐藏层的简单神经网络,通过训练该网络得到隐藏层的权重,这些权重就是希望学习的词向量。Skip-gram 算法已经集成在 word2vec 开源包中,可直接调用该软件包训练词向量。将由 Skip-gram 模型训练得到的词向量存储在一个词嵌入矩阵 $E \in \mathbb{R}^{n \times |V|}$ 中,其中, $|V|$ 表示词汇表的大小, n 为词向量维度。假设一个语句 X 包含有 j 个词,则此语句可以表示为 $X_{[1:j]}$,每个词语在词嵌入矩阵 E 中都有一个唯一的用于检索其对应词向量的索引 k ,句子中第 i 个词的词向量用 x_i 表示:

$$x_i = Eb_k \quad (1)$$

其中, b_k 是一个维度为词表大小,值为 0 或 1 的二值向量,除了第 k 个索引之外的所有位置都是 0。则一个分词后的中文句子可以用矩阵 X 表示,计算公式如下:

$$X = (x_1, x_2, \dots, x_j) \quad (2)$$

在训练本文模型的过程中还将对词向量和其他模型参数进行微调,以期能够达到最佳分类效果。

2 网络模型

为了有效提高新闻文本分类的准确率,本文提出一种引入注意力机制的 CNLSTM 模型用于中文新闻文本分类,其结构如图 1 所示。模型主要包括 4 个部分:第 1 部分是短语特征序列提取操作,主要使用一维卷积对词向量提取特征,利用滑动窗口计算前后词对当前词的影响,生成短语特征表示;第 2 部分是新闻文本特征提取操作,该部分使用 NLSTM 处理短语特征序列,逐步合成文本的向量特征表示;第 3 部分采用注意力机制计算短语的重要性分布,生成含有注意力概率分布的文本特征表示;第 4 部分是分类器,主要由 dropout 技术防止过拟合,用 Softmax 分类器预测文本类别。

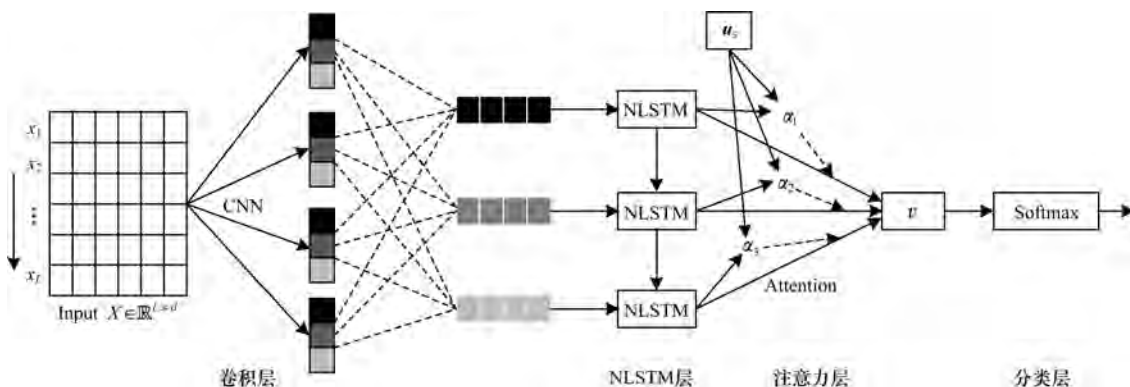


图 1 CNLSTM 模型结构

2.1 短语特征序列提取

本文采用一维卷积核在文本的不同位置滑动来提取词语的上下文信息,生成短语的特征表示。用 $x_i \in \mathbb{R}^d$ 表示句子中第 i 个词语的 d 维词向量表示, $X \in \mathbb{R}^{L \times d}$ 代表输入句子, 其中 L 是句子的长度。用 $k \in \mathbb{R}^{m \times d}$ 表示卷积操作所采用的卷积核向量, m 为卷积核窗口大小。则在句子中的第 j 个位置上, 可以得到由 m 个连续词向量组成的一个窗口矩阵 w_j , 其表示如下:

$$w_j = [x_j, x_{j+1}, \dots, x_{j+m-1}] \quad (3)$$

其中, 逗号表示行向量连接。卷积核 k 以有效的方式在每个位置处与窗口向量 (m -gram) 进行卷积以生成输入文本的特征映射 $F \in \mathbb{R}^{L-m+1}$, 其原理如图 2 所示。

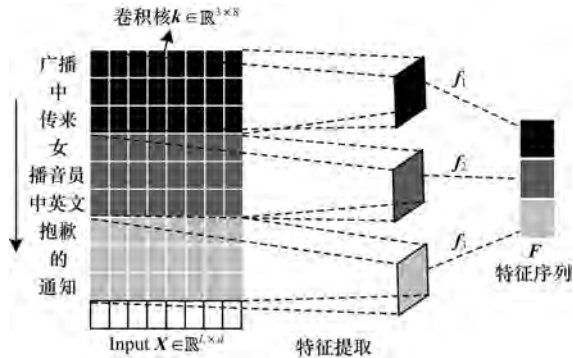


图2 短语特征提取

在图2中, 输入的句子为“广播/中/传来/女/播音员/中英文/抱歉的/通知”, 每个字用一个8维向量表示, 卷积核的窗口大小为3, 不同颜色的网格代表同一卷积核在输入句子的不同位置进行特征提取的过程。由于是一维卷积, 图中向下的箭头表示卷积过程中卷积核的移动方向自上而下, 即只在行方向上进行卷积。此外, 为了方便说明卷积过程, 图2所示的卷积操作的步长 stride 设为3。图2中 f_1, f_2, f_3 代表提取获得的短语特征, 例如: f_3 代表短语“抱歉的通知”的特征。 f_1, f_2, f_3 连接后的特征向量 F 即是整个句子的特征表示。特征映射 F 的每个元素 f_j 计算方法如下:

$$f_j = g(w_j \odot k + b) \quad (4)$$

其中, \odot 是矩阵的元素相乘, b 是偏置项, g 是激活函数, 本文选用修正线性单元 ReLU^[13]。在模型中使用多个卷积核以产生多个特征映射。用 n 个具有相同长度的卷积核, 所生成的 n 个特征映射可以被重新排列为特征表示:

$$T = [F_1; F_2; \dots; F_n] \quad (5)$$

其中, 分号表示列向量连接, F_i 是用第 i 个卷积核在整个输入数据上卷积生成的特征映射, $T \in \mathbb{R}^{(L-m+1) \times n}$ 中的每一行 T_j 是用 n 个卷积核在句子中的同一位置 j 处进行卷积操作产生的新的特征表示。然后将这个新的、连续的高阶特征表示 T 传送至下

文描述的 NLSTM 中。

通过卷积操作提取短语特征之后, 会进一步执行池化操作, 但是文本分类是利用现有信息进行预测, 强调特征序列的连续性, 而池化操作会破坏特征序列的连续性。因此, 本文在 CNN 卷积层之后直接采用 NLSTM 模型。

2.2 文本特征提取

虽然通过 CNN 提取到了短语的特征, 但在卷积的过程中并没有考虑短语之间的相互联系。此外, 由于中文文本中通常含有倒装、前置等复杂的表达形式, 在文本分类过程中可能需要用到以前的某些历史信息。LSTM 作为一种经典的改进结构, 它采用门结构调节信息流动, 通过存储单元存储历史信息, 因此被广泛应用于各种应用中, 且表现突出。

基于 LSTM 的特点, 本文采用文献[5]提出的 NLSTM 结构。NLSTM 通过嵌套增加 LSTM 的深度, 即 NLSTM 中记忆单元的值 c_t 是通过一个 LSTM 计算而来, 其结构如图3所示。

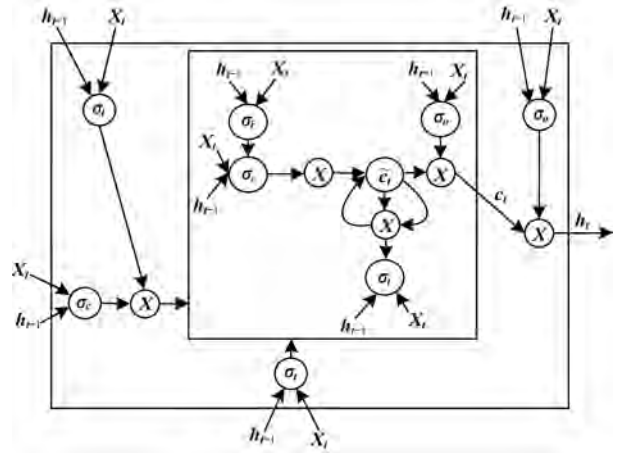


图3 NLSTM 记忆单元结构

在 LSTM 中, 各状态和各门的更新方式如下:

$$i_t = \sigma_i(x_t W_{xi} + h_{t-1} W_{hi} + b_i) \quad (6)$$

$$f_t = \sigma_f(x_t W_{xf} + h_{t-1} W_{hf} + b_f) \quad (7)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \sigma_c(x_t W_{xc} + h_{t-1} W_{hc} + b_c) \quad (8)$$

$$o_t = \sigma_o(x_t W_{xo} + h_{t-1} W_{ho} + b_o) \quad (9)$$

$$h_t = o_t \odot \sigma_h(c_t) \quad (10)$$

而 NLSTM 通过一个已经学习的状态函数式(11)来替代 LSTM 中计算 c_t 的加运算。

$$c_t = m_t(f_t \odot c_{t-1}, i_t \odot g_t) \quad (11)$$

状态函数 m_t 被称为 t 时刻的内部记忆, 将这个状态函数作为另一个 LSTM 的记忆单元, 产生一个 NLSTM 网络。在 NLSTM 中, 内部记忆函数的输入和隐层状态就变为:

$$\tilde{h}_{t-1} = f_t \odot c_{t-1} \quad (12)$$

$$\tilde{x}_t = i_t \odot \sigma_c(x_t W_{xc} + h_{t-1} W_{hc} + b_c) \quad (13)$$

则内部 LSTM 的状态和门控信息的更新过程为:

$$\tilde{i}_t = \sigma_i(\tilde{x}_t \tilde{W}_{xi} + \tilde{h}_{t-1} \tilde{W}_{hi} + \tilde{b}_i) \quad (14)$$

$$\tilde{f}_t = \sigma_f(\tilde{x}_t \tilde{W}_{xf} + \tilde{h}_{t-1} \tilde{W}_{hf} + \tilde{b}_f) \quad (15)$$

$$\tilde{c}_t = \tilde{f}_t \odot \tilde{c}_{t-1} + \tilde{i}_t \odot \sigma_c(\tilde{x}_t \tilde{W}_{xc} + \tilde{h}_{t-1} \tilde{W}_{hc} + \tilde{b}_c) \quad (16)$$

$$\tilde{o}_t = \sigma_o(\tilde{x}_t \tilde{W}_{xo} + \tilde{h}_{t-1} \tilde{W}_{ho} + \tilde{b}_o) \quad (17)$$

$$\tilde{h}_t = \tilde{o}_t \odot \tilde{\sigma}_h(\tilde{c}_t) \quad (18)$$

外部 LSTM 的记忆单元状态信息更新为:

$$c_t = \tilde{h}_t \quad (19)$$

由此可见,在 NLSTM 中,LSTM 通过标准的门控结构选择性地访问内部存储单元,这一关键特征使得 NLSTM 比传统的堆栈式 LSTM 实现更有效的时间层级。因为在访问内部记忆时,NLSTM 拥有更高的访问自由度,从而能处理更长时间规模的历史信息。且实验结果证明,在参数量相同的情况下,NLSTM 表现优于堆栈式 LSTM 和单层 LSTM,与堆栈式的多层 LSTM 相比,NLSTM 的内部记忆单元能够学习更长期的依赖关系。

2.3 注意力分布计算

注意力机制是一种资源分配制度,它通过模拟人脑注意力的特点,对重要的信息给予较多的注意力,而对其他内容分配较少的注意力。目前注意力机制已被应用于各领域中,如文献[14]将注意力机制应用于 RNN 模型进行图像分类;文献[15]将注意力机制应用于机器翻译任务中,提出基于注意力机制的 encode—decode 模型,大幅提高了翻译效果;文献[16]基于注意力机制进行微博情感分析,也取得了较好的结果。在自然语言处理中引入注意力机制计算注意力的分布概率,突出输入各部分对输出的影响程度,可以达到优化传统模型的效果。本文在 NLSTM 模块之后引入注意力机制,以生成含有注意力概率分布的文本语义特征表示,由此突出输入文本中不同短语特征对文本类别的影响作用,提高模型分类的准确度。引入注意力机制后,文本的特征表示计算过程如下:

$$u_t = \tanh(W_s h_t + b_s) \quad (20)$$

$$\alpha_t = \text{Softmax}(u_t^T, u_s) \quad (21)$$

$$v = \sum_i \alpha_i h_i \quad (22)$$

其中, h_t 是由 NLSTM 学习得到的 t 时刻的特征表示, u_t 为 h_t 通过一个简单神经网络层得到的隐层表示, u_s 是一个随机初始化的上下文向量,可视为对输入的一种语义表示, α_t 为 u_t 通过 Softmax 函数归一化得到的重要性权重, v 即是最终文本信息的特征向量。

2.4 Softmax 分类

为防止模型在训练时出现过拟合现象,本文引

入 dropout 技术,并采用 Softmax 分类器对所获得的文本特征进行多分类处理:

$$y = \text{Softmax}(v) \quad (23)$$

其中, y 是一个维度为类别数量大小的向量,其每一维都是一个[0,1]范围内的数字,代表该文本属于某个类别的概率。 n 是可能的类别个数,则输入句子的类别为:

$$\hat{y} = \text{argmax}(y) \quad (24)$$

在本文模型的训练过程中,通过最小化输出类别与句子真实类别之间的交叉熵误差训练整个模型。给定训练样本 x 和样本的真实类别 l ,则其交叉熵误差为:

$$E(x, l; \theta) = - \sum_{i=1}^n l_i \ln(y_i; \theta) \quad (25)$$

其中, θ 为模型的参数, l_i 指实际的类别标签向量中的第 i 个值, y_i 为 Softmax 的输出向量 y 的第 i 个值。对于所获得的误差,最后取其平均即是该模型的损失函数。此外,本文还对其进行 L_2 正则化,则最终的目标函数为:

$$J(\theta) = \frac{1}{N} E(x, l; \theta) + \frac{\lambda}{2} \|\theta\|^2 \quad (26)$$

其中, λ 是 L_2 正则项系数, N 为训练样本大小。在训练模型的时候可以采用 Adam 方法来最小化目标函数 $J(\theta)$,它主要利用梯度的一阶矩估计和二阶矩估计动态调整每个参数的学习率。Adam 的优点主要在于经过偏置校正后,每一次迭代学习率都有个确定范围,使得参数变化比较平稳。

综上,本文进行文本分类时所用算法描述如下,其中, *TrainingData* 为训练数据集(有标签的中文语料库), θ 为模型参数集, J 为优化目标函数, J_{sum} 为训练集误差和, X 为训练集中的一个新闻文本, l 为训练样本的真实类别, L 为词数, d 为词向量维度, n 个卷积核的大小均为 k ,*epochNum* 为总迭代次数。

算法 1 CNLSTM + Attention 模型训练算法

输入 *TrainingData*

输出 权值确定的多层模型

1. 初始化参数 θ ,用 Word2vec 训练获得词向量;

While 迭代次数 < *epochNum*

2. $J \leftarrow 0$; $J_{sum} \leftarrow 0$;

3. For $X \in \text{TrainingData}$

4. $E(X, l; \theta) \leftarrow 0$;

5. 输入 X 得出模型产生的 X 的类别概率向量 y :

$y = \text{model}(X; \theta)$

6. 计算训练样本 X 的误差:

$$E(X, l; \theta) = - \sum_{i=1}^n l_i \log(y_i; \theta)$$

7. 计算训练样本集损失:

$$J_{sum} \leftarrow J_{sum} + E(X, l; \theta)$$

8. End for

$$9. J \leftarrow J_{sum} + \frac{\lambda}{2} \|\theta\|^2$$

10. 用 Adam 计算梯度 $\frac{\partial J}{\partial \theta}$ 并更新 θ ;

11. End while

经过上述训练步骤后,本文网络模型即可用于对未知类别的测试集进行分类处理,将测试数据集中的新闻样本 X 输入到模型中,该模型的输出即为 X 所属类别的概率向量 y ,取 y 中最大概率对应的类别为测试样本 X 的类别。其具体算法如下,其中, $TestData$ 为测试数据集, \hat{y} 为所求测试样本的类别, X 为测试数据集中的一条新闻文本, L 为词数, d 为词向量维度。

算法2 测试样本分类算法

输入 $TestData$

1. 初始化:对测试集进行分词,然后用 Word2vec 求得词向量。

2. For $X \in TestData$

3. 向模型中输入 X ,得出 X 的类别概率向量 y :

$y = \text{model}(X; l)$

4. 输出:该样本的类别

$\hat{y} = \text{argmax}(y)$

5. End for

3 实验与结果分析

3.1 实验数据

本文使用了3个数据集进行文本分类,以测试 Attention-based CNLSTM 模型的分类效果,包括 THUCNews 新闻数据集(下载地址: <http://thuctc.thunlp.org/message>)、Sogou 新闻语料库(下载地址: <http://www.sogou.com/labs/resource/tce.php>)以及复旦新闻语料库(下载地址: <https://dvn.fudan.edu.cn/dataverse.xhtml>)。

THUCNews 是由清华大学自然语言处理实验室推出的新闻文本数据集,包含74万篇新闻文档,分为14个候选新闻类别。本文使用其中的10个类别,每个类别包含5000个训练样本,500个验证集样本和1000条测试样本。

Sogou 语料库是搜狗实验室提供的全网新闻数据,该数据集来自2012年6月—7月新浪、网易、腾讯以及凤凰资讯等若干个新闻站点。由于完整的实验数据量过于庞大,本文使用其中的9个类别,每个

类别包含2000个训练样本,500个验证集样本和500条测试样本。

复旦大学新闻语料库由该校李荣路老师整理并提供,分为20个新闻话题类别,共包括9000多个文档。但由于该数据集中某些类别数量过少,因此本文实验不使用文本数量低于1000的类别文本。本文使用了5个类别,每个类别包含1500个训练样本,500个验证集,500条测试样本。

3.2 模型训练及实验设置

本文采用 mini-batch 梯度下降法进行模型的训练。网络模型为1层卷积层、1层 NLSTM 层和1层注意力层,参数设置如表1所示。

表1 参数设置

参数	设置
卷积核大小	3,4,5,6
卷积核数量	256
词向量维度	64
丢弃率 P	0.5
mini-batch	128
句子长度	1000
嵌套深度 depth	2,3,4,5,6
学习率	0.01
正则项因子	0.001

3.3 实验结果

为了测试模型的有效性,本文选择了多个目前在中文文本分类任务上应用广泛且效果较好的网络结构作为 Baseline 模型进行比较。比较模型包括:单层 CNN、单层 LSTM、单层 NLSTM、CNN + NLSTM、Attention-based LSTM^[17]、BiLSTM^[18]、Attention-based BiLSTM^[19]、BiLSTM + max-pooling^[20]。本文也引用在相同数据集上已有的代表性工作作为比较,其中:文献[21]提出一种 VDCNN 与 LSTM 相结合的混合模型,该混合模型在搜狗和复旦数据集上取得了较好效果;文献[22]的方法对整个文本的词向量进行多种池化,最后将多种池化的特征作为一个整体输入到 Softmax 回归模型中得到文本的类别信息,该方法在复旦大学的数据集上取得了较好效果。具体对比结果如表2所示。

表2 模型精确度对比

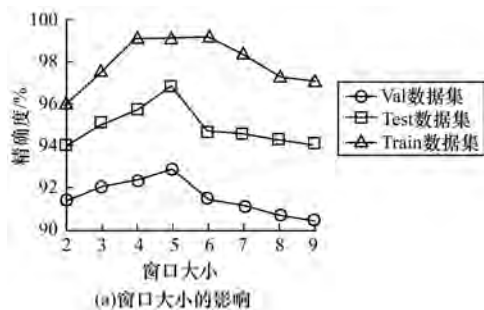
%

模型	THUCNews 数据集	复旦数据集	Sogou 数据集
CNN	92.37	90.30	91.34
LSTM	90.86	89.57	92.18
NLSTM	93.39	91.84	94.53
CNN + NLSTM	95.64	94.21	96.77
Attention-based LSTM ^[17]	91.16	92.00	95.33
BiLSTM ^[18]	80.68	83.21	89.40
Attention-based BiLSTM ^[19]	86.32	94.01	94.34
BiLSTM + max-pooling ^[20]	85.16	82.81	90.24
VDCNN + LSTM ^[21]		93.10	98.96
多种池化 ^[22]			94.73
本文方法	96.87	95.43	97.58

从表 2 可以看出,NLSTM 的分类准确率均高于 LSTM,由此可以说明 NLSTM 的确能够更有效地学习到文本中的长期依赖关系。从 CNN 与 CNN + NLSTM 的对比结果可以看出,在 CNN 中引入 NLSTM 网络层,使得分类准确度得到显著提升。最终注意力机制的引入更进一步提升了模型的性能。通过表 2 还可以看出,在 THUCNews 数据集和复旦数据集上,本文模型优于其他的比较模型取得了最佳的效果。与单层 CNN 和单层 NLSTM 的比较表明,本文混合模型的优点在于不但能够更好学习到句子的全局特征表示,而且能更好地学习具有注意力分布的更高级别表示的长期依赖关系。特别是通过与 Attention-based LSTM 的比较,更加突出显示了本文模型的优越性。虽然在 Sogou 数据集上,本文模型取得的效果不是最佳,低于 VDCNN + LSTM,但与其他模型相比,本文模型的分类准确率高,且接近 VDCNN + LSTM 的准确率。

3.4 模型分析

本文以 THUCNews 新闻数据集为例进行 10 分



类验证,考察了卷积核窗口大小、NLSTM 嵌套深度 2 个因素对模型精确度的影响。

在本文模型的卷积层中,卷积核用于捕获局部 n -gram 特征,即短语特征。因此不同大小的卷积核可以捕获不同的 n -gram 特征。本文在其他参数保持一致的情况下,以不同卷积核大小做实例验证观察模型预测精确度的变化,结果如图 4(a)所示。从图中可以得出:当窗口大小低于 5 时,模型的精确度随着卷积核窗口的增大而逐渐增大;但窗口大小大于 5 时,模型的精确度反而随窗口增大而减小。由此可以得出:对于 THUCNews 数据集而言,窗口大小为 5 的卷积核能够捕捉到更加充分、准确的短语特征使得本文模型达到最佳效果。此外,由于随着 NLSTM 嵌套深度的增加,整个模型的复杂度也会增加,因此需要选择一个合适的嵌套深度,使得模型在该深度下可以取得最佳效果。图 4(b)是在 THUCNews 数据集上使用不同深度得到的本文模型精确度的变化结果,可以看出对该数据集而言,最佳嵌套深度为 5。

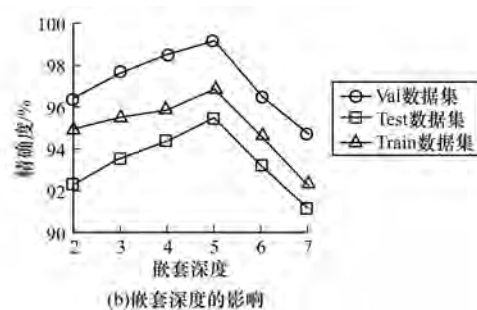


图 4 各参数对实验效果的影响

4 结束语

本文利用 CNLSTM 模型对词向量化后的文本进行分类处理。CNLSTM 模型将 CNN 和 NLSTM 相结合,利用卷积层学习短语级别的特征表示,这些高层的短语特征表示序列被输入到 NLSTM 网络中进一步学习短语间的长短依赖关系,得出整个输入文本的特征表示。在此基础上,引入注意力机制捕获输入文本中的重要信息,减少特征提取过程中的信息丢失和信息冗余问题。在 3 个公开新闻数据集上测试模型性能,结果表明,该混合模型在特征提取过程中,具备保留历史信息且利用前后文信息的能力,弥补了简单卷积层的不足,从而有效提高了文本分类的准确率。下一步将优化 CNLSTM 模型,在提高模型分类准确度的情况下同时提升其运行速度。

参考文献

[1] 陈钊,徐睿峰,桂林,等.结合卷积神经网络和词语情感序列特征的中文情感分析[J].中文信息学报,2015,29(6):172-178.

[2] 夏从零,钱涛,姬东鸿.基于事件卷积特征的新闻文本分类[J].计算机应用研究,2017,34(4):991-994.

[3] MIKOLOV T, SUTSKEVER I, CHEN Kai, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of Advances in Neural Information Processing Systems. [S. l.]: Neural Information Processing Systems Foundation, Inc., 2013: 3111-3119.

[4] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[5] MONIZ J R A, KRUEGER D. Nested LSTM [EB/OL]. [2018-04-01]. <https://arxiv.org/pdf/1801.10308.pdf>.

[6] AGARWAL B, MITTAL N. Text classification using machine learning methods; a survey [C]//Proceedings of the 2nd International Conference on Soft Computing for Problem Solving. Berlin, Germany: Springer, 2014: 701-709.

[7] 李荣艳,金鑫,王春辉,等.一种新的中文文本分类算法[J].北京师范大学学报(自然科学版),2006,42(5): 510-505.

(下转第 314 页)

(上接第 308 页)

- [8] PENG Fuchun, SCHUURMANS D. Combining naive Bayes and n-gram language models for text classification [C]//Proceedings of European Conference on Information Retrieval. Berlin, Germany: Springer, 2003: 335-350.
- [9] 翟林,刘亚军. 支持向量机的中文文本分类研究[J]. 计算机与数字工程, 2005, 33(3): 21-23, 45.
- [10] KALCHBRENNER N, GREFFENSTETTE E, BLUNSON P. A convolutional neural network for modelling sentences [EB/OL]. [2018-04-01]. <https://arxiv.org/pdf/1404.2188.pdf>.
- [11] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model [J]. Journal of Machine Learning Research, 2003, 3: 1137-1155.
- [12] 谢逸, 饶文碧, 段鹏飞, 等. 基于 CNN 和 LSTM 混合模型的中文词性标注[J]. 武汉大学学报(理学版), 2017, 63(3): 246-250.
- [13] NAIR V, HINTON G E. Rectified linear units improve restricted Boltzmann machines[C]//Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel: [s. n.], 2010: 807-814.
- [14] MNIH V, HEES N, GRAVES A, et al. Recurrent models of visual attention[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014: 2204-2212.
- [15] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [EB/OL]. [2018-04-01]. <https://arxiv.org/pdf/1409.0473.pdf>.
- [16] 周瑛, 刘越, 蔡俊. 基于注意力机制的微博情感分析[J]. 情报理论与实践, 2018, 41(3): 85-94.
- [17] 张冲. 基于 Attention-Based LSTM 模型的文本分类技术的研究[D]. 南京: 南京大学, 2016.
- [18] 黄磊, 杜昌顺. 基于递归神经网络的文本分类研究[J]. 北京化工大学学报(自然科学版), 2017, 44(1): 98-104.
- [19] 胡朝举, 梁宁. 基于深层注意力的 LSTM 的特定主体情感分析[J]. 计算机应用研究, 2019, 36(5): 10-15.
- [20] LAI Siwei, XU Liheng, LIU Kang. Recurrent convolutional neural networks for text classification[C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Publications, 2015: 2267-2273.
- [21] 彭玉青, 宋初柏, 闫倩, 等. 基于 VDCNN 与 LSTM 混合模型的中文文本分类研究[J]. 计算机工程, 2018, 44(11): 190-196.
- [22] 阳馨, 蒋伟, 刘晓玲. 基于多种特征池化的中文文本分类算法[J]. 四川大学学报(自然科学版), 2017, 54(2): 287-292.

编辑 刘盛龄