

doi:10.11835/j.issn.1000-582X.2023.05.012

基于改进长短时记忆网络的文本分类方法

李建平, 陈海鸥

(东北石油大学 计算机与信息技术学院, 黑龙江 大庆 163318)

摘要:针对传统长短时记忆网络(long short-term memory,LSTM)在文本分类中无法自动选取最重要潜在语义因素的问题,提出一种改进的 LSTM 模型。首先,将传统 LSTM 的运算关系拓展为双向模式,使网络充分记忆输入特征词的前后关联关系;然后在输出层前面增加池化层,以便更好选择找到最重要的潜在语义因素。互联网电影资料库评论数据实验结果表明,该模型优于传统长短时记忆神经网络以及其他同类模型,揭示了改进方案对提高文本分类准确率是有效的。

关键词:自然语言处理;文本分类;循环神经网络;长短时记忆神经网络

中图分类号:TP391

文献标志码:A

文章编号:1000-582X(2023)05-111-08

Text classification method based on improved long-short term memory network

LI Jianping, CHEN Haiou

(School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, Heilongjiang, P. R.China)

Abstract: Traditional long-short term memory network (LSTM) cannot automatically select the most important latent semantic factors in text categorization. To solve the problem, this paper proposes an improved LSTM model. First, the traditional LSTM operation relationship is extended to the bidirectional mode, so that the network fully remembers the context of the input feature words. Then, the pooling layer is added in front of the output layer to better select the most important latent semantic factors. The experiment on the Internet Movie Database review data show that the model is superior to the traditional long-short term memory neural network and other similar models, revealing that the improved scheme proposed in this paper can improve the accuracy of text classification.

Keywords: natural language processing; text classification; recurrent neural network; long-short term memory neural network

随着互联网行业蓬勃发展,大数据进入人们生活。在大数据时代,数据有着多样的表达方式,如最传统的文本格式,能直观获取信息的图片格式,深受喜爱的短视频形式,节约时间成本的语音格式等。其中,文本作为信息的一种重要承载方式,相对于其他信息载体,具有容量小,方便存储和管理的特点。但是,文本无法

收稿日期:2021-08-11

基金项目:国家自然科学基金资助项目(61702093)。

Supported by National Natural Science Foundation of China(61702093).

作者简介:李建平(1976—),男,博士,教授,主要从事网络信息安全、自然语言处理方向研究,(E-mail) ljp@nepu.edu.cn。

像图像、视频或语音等直观表示其内在含义,更多依靠人工阅读理解才能得到含义,然而在如今海量的文本信息中,人工实现对文本内容定义十分困难。因此,在海量文本信息中,快速提取人们所需尤为重要^[1]。文本分类成为许多人研究的内容。特别是短文本分类随着社交媒体的发展,应用深度学习进行情感分析^[2-3]越来越重要。

文本分类中一个重要问题是特征表示^[4],人们通常使用词袋模型(Bag-of-Words),其中一元模型、二元模型、 n 元模型或一些精心设计的模式作为选择特征,也有基于文本的特质分类模式^[5],或给予被噎死的深度特征加权,多标签文本分类^[6]。此外,一些特征选择方法,如频率、MI、PLSA^[7]、LDA^[8]、FSDBN^[9],使用这些方法进行特征选取,提高选取特征的分类效率。以往的特征选取方法通常会忽略文本中上下文或是文本中词序关联,并没有分类中使用的文本词汇的语义选取效果更佳。后来,使用预训练嵌入:词向量^[10],特别是近年来,神经网络的高速发展,带给自然语言处理工作新的发展方向。首先是单词嵌入,这种将词嵌入到一整层,从而解决数据分散问题的方法,受到研究者喜爱,结合预训练嵌入与单词嵌入,获得提升分类效果的句法效率规则。在此基础上,有人提出一些基于构图的文本语义表示方法。近年来,计算力大幅度提升,特别是超级计算机,量子计算机的开发与应用,许多人开始认真对待深度学习。将深度学习与自然语言处理相结合。使用卷积网络进行文本分类^[11]。在构建语言模型方面,开始采用循环神经网络(recurrent neural network, RNN)进行句法解析,将RNN建立起深层结构,成为典型深度学习学习模型^[12]。但是循环神经网络在进行实验的过程中会出现梯度爆炸和梯度消失的问题^[13],研究者将 Jürgen Schmidhuber 提出的长短时记忆网络^[14],将长短时记忆网络与本文分类相结合。长短时记忆网络在长文本的分类中起到了十分重要的作用。虽然长短时记忆网络对文本分类的成功率有一定提升,但利用 LSTM 对句子进行建模存在一个问题:无法编码从后到前的信息。

针对以上传统 LSTM 分类模型的不足,笔者提出一种带池化层的双向 LSTM 模型(Bi-LSTM-pooling)。首先将传统 LSTM 扩展为双向 LSTM(即 BiLSTM)。与传统 LSTM 相比,他可以更加准确地捕捉文本序列中上下文之间的关联关系,以便在学习单词表示时尽可能准确地捕获上下文信息。其次,在 BiLSTM 的输出层之前引入池化层。实验采用最大池化层对文本提取的特征进行判断,分辨特征的分类效果,从而提出文本中对分类更关键成分。该模型可融合循环神经网络和卷积神经网络的优点。互联网电影资料库(IMDB)评论数据实验结果表明,该模型的分类性能优于传统长短时记忆网络,从而验证了改进措施的有效性。

1 传统 LSTM 模型

1.1 循环网络

针对特定任务文本分类的递归神经网络,神经模型的主要作用是将变长文本表示为固定长度的向量。循环神经网络(RNN)处理任意长度的序列一般使用递归方法,使用转换函数与输入序列的内部隐藏状态向量 h_t 相结合,起到处理作用,循环神经网络的结构如图 1 所示。

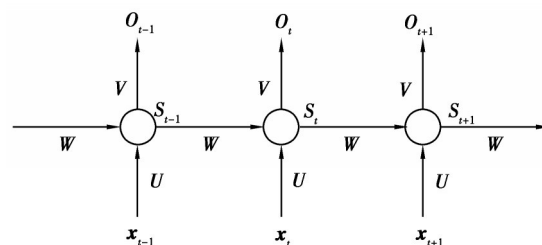


图1 循环神经网络 RNN 结构图

Fig.1 RNN structure diagram of recurrent neural network

x_t 是一个 n 维向量,每一个 x_t 代表一个词向量, s_t 是时间 t 处的记忆

$$s_t = Ux_t + Wh_{t-1} + b, \quad (1)$$

\mathbf{h}_t 代表 t 时刻的隐藏状态,一般使用非线性的函数作为转换函数,如 \tanh 等函数, W 为隐藏层到隐藏层的权重

$$\mathbf{h}_t = f(U\mathbf{x}_t + W\mathbf{h}_{t-1} + b), \quad (2)$$

其中: o_t 代表 t 时刻的输出; g 一般为 softmax 函数; V 为隐藏层到输出层的权重

$$o_t = g(V\mathbf{h}_t + c), \quad (3)$$

利用 \mathbf{x}_t 和 \mathbf{h}_{t-1} 的变换,将状态转换函数 f 作为元素非线性的组成,建模序列使用 RNN 将输入序列映射到固定大小的向量,将该向量馈送到一个分类的 softmax 层。然而具有这种形式的转换函数的 RNN 有一个问题,在训练过程中,梯度向量的分量可以在长序列上呈指数增长或衰减。这一爆炸或消失的问题 RNN 模型很难在序列中学习长距离相关。

1.2 长短时记忆网络

Hochreiter 和 Schmidhuber 提出了长期短期记忆网络^[12]来解决长期依赖关系的问题, LSTM 在其内部保留了一个独立的记忆单元,仅在必要时才显示其内容。在每个时间步骤 t 将 LSTM 单位定义为 \mathbb{R}^d 中的向量集合:遗忘门、输入门、细胞状态 $C^{(t)}$ 、输出门和隐藏状态 $\mathbf{h}^{(t)}$ 。 d 是 LSTM 装置的数量。LSTM 转换过程如下

1)更新遗忘门输出

$$f^{(t)} = \sigma(W_f \mathbf{h}^{(t-1)} + U_f \mathbf{x}_t^{(t)} + b_f), \quad (4)$$

遗忘门是控制是否遗忘,以一定的概率控制是否遗忘上一层的隐藏细胞状态。

2)更新输入门输出

$$i^{(t)} = \sigma(W_i \mathbf{h}^{(t-1)} + U_i \mathbf{x}_t^{(t)} + b_i), \quad (5)$$

$$a^{(t)} = \tanh(W_a \mathbf{h}^{(t-1)} + U_a \mathbf{x}_t^{(t)} + b_a), \quad (6)$$

输入门是在遗忘门进行遗忘后,进行补充的部分,主要负责处理当前序列位置的输入,由 2 部分组成:用 sigmoid 激活函数,输出为 i_t ;用 tanh 激活函数,输出为 a_t ,结果进行相乘来更新细胞状态。

3)更新细胞状态

$$C^{(t)} = C^{(t-1)} \odot f^{(t)} + i^{(t)} \odot a^{(t)}, \quad (7)$$

细胞状态存储信息,遗忘门、输入门的结果作用于细胞状态, \odot 表示元素乘法。

4)更新输出门输出

$$o^{(t)} = \sigma(W_o \mathbf{h}^{(t-1)} + U_o \mathbf{x}_t^{(t)} + b_o), \quad (8)$$

$$\mathbf{h}^{(t)} = o^{(t)} \odot \tanh(C^{(t)}), \quad (9)$$

输出门控制哪些信息需要作为输出。

5)更新当前序列索引预测输出

$$\hat{\mathbf{y}}^{(t)} = \sigma(V\mathbf{h}^{(t)} + c), \quad (10)$$

但利用 LSTM 对句子进行建模还存在无法编码从后到前的信息。

2 改进的 LSTM 模型

2.1 双向 LSTM 模型

为解决无法编码从后到前的信息问题,使用 Bi-LSTM 模型进行改进,此模型可以更好捕捉双向的语义依赖。

在改进中,改进模型的输入是文档,是单词 $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n$ 的序列,输出是分类结果。使用 $p(k|D, \theta)$ 来表示文档为 k 类的概率, θ 是模型中的参数。结合一个词和它的上下文来表示一个词。上下文有助于获得更精确的词义,在模型中,使用双向长短时记忆网络来对文本的整体进行信息的提取。将 $\mathbf{c}_l(\mathbf{w}_i)$ 定义为单词 \mathbf{w}_i 的左上下文,将 $\mathbf{c}_r(\mathbf{w}_i)$ 定义为单词 \mathbf{w}_i 的右上下文。 $\mathbf{c}_l(\mathbf{w}_i)$ 和 \mathbf{w}_i 都是有 $|c|$ 实值向量。单词 \mathbf{w}_i 的左侧上下文 $\mathbf{c}_l(\mathbf{w}_i)$ 使用式(11)计算,其中: $\mathbf{e}(\mathbf{w}_{i-1})$ 是单词 \mathbf{w}_{i-1} 的嵌入词,它是一个实值向量; $\mathbf{c}_l(\mathbf{w}_{i-1})$ 是 \mathbf{w}_{i-1} 的左侧上下文。文中 \mathbf{w}_i 的左侧上下文使用相同的共享参数 $\mathbf{c}_l(\mathbf{w}_i)$ 。 $\mathbf{W}^{(l)}$ 是将隐藏层转换为下一个隐藏层的矩阵。 $\mathbf{W}^{(sl)}$ 是一个矩

阵,用于结合 w_i 和 w_{i+1} 的左上下文。 f 是一个激活函数。 $c_r(w_i)$ 的方式计算,如式(12)所示。文本最后一个单词的右上下文共享参数 $c_r(w_n)$ 。

$$c_l(w_i) = f(W^{(l)}c_l(w_{i-1}) + W^{(sl)}e(w_{i-1})), \tag{11}$$

$$c_r(w_i) = f(W^{(r)}c_r(w_{i+1}) + W^{(sr)}e(w_{i+1})). \tag{12}$$

通过 $c_l(w_i)$ 和 $c_r(w_i)$ 来提取文本中的信息,定义了 w 的表示。在式(13)中,它是左侧上下文向量 $c_l(w_i)$ 、单词嵌入 $e(w_i)$ 和右侧上下文向量 $c_r(w_i)$ 的串联。通过以上方式,可以减少 w_i 所带来的歧义。

$$x_i = [c_l(w_i); e(w_i); c_r(w_i)]. \tag{13}$$

循环结构可以在文本的前向扫描中获得所有 c_l ,在文本的后向扫描中获得 c_r 。时间复杂度为 $O(n)$ 。在获得单词 w_i 的表示 x_i 之后,将线性变换与 \tanh 激活函数一起应用到 x_i ,并将结果发送到下一层。

$$i_i^{(2)} = \tanh(W^{(2)}x_i + b^{(2)}), \tag{14}$$

$y_i^{(2)}$ 是一个潜在的语义向量,在这个向量中,每一个语义因子都会被分析,以确定最有用的表示文本的因素。

2.2 双向 LSTM 加池化层模型

模型中的卷积神经网络是从卷积神经网络的角度来表示文本,前面提到的循环结构是卷积层。当计算出所有单词的表示形式时,将应用最大化池层。

$$y^{(2)} = \max_{i=1}^n y_i^{(2)}, \tag{15}$$

其中, \max 函数是一个按元素排序的函数。 $y^{(3)}$ 的 k 元素是 $y_i^{(2)}$ 的 k 元素中最大的元素。

通过池层将不同长度的文本转换为固定长度的向量。当然,在深度学习中也有其他类型的池层,例如平均池层等,但其对于获取文本中对分类有效成分的作用并没有很突出。因此,笔者使用最大池层的原因是其能够在文本中选取对于文本分类更有效果的词。池层利用双向长短时记忆网络的输出作为自己的输入。池层的时间复杂性是 $O(n)$ 。总体模型是循环结构和最大池层,最后模型的时间复杂性仍然是 $O(n)$ 。

模型的最后一部分是输出层。类似于传统的神经网络。它被定义为

$$y^{(4)} = W^{(4)}y^{(3)} + b^{(4)}, \tag{16}$$

最后,将 SoftMax 函数应用于 $y^{(4)}$,将输出数字转换为概率

$$p_i = \frac{\exp(y_i^{(4)})}{\sum_{k=1}^n \exp(y_k^{(4)})}. \tag{17}$$

综上所述,提出的改进 LSTM 模型如图 2 所示。

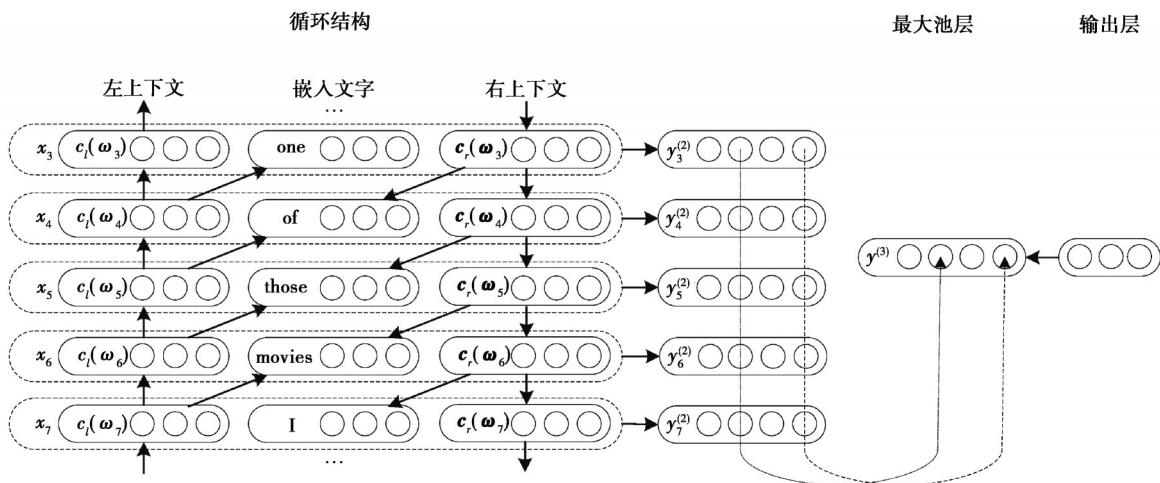


图 2 改进 LSTM 结构图

Fig.2 Improved LSTM structure diagram

2.3 文本分类实施方案

在实施分类之前,首先要将数据集中的文本对照词汇表进行数据预处理,并且训练词向量。并带入模型进行训练。训练网络参数将所有要训练的参数定义为 θ

$$\theta = \{ E, \mathbf{b}^{(2)}, \mathbf{b}^{(4)}, \mathbf{c}_l(\mathbf{w}_1) + \mathbf{c}_r(\mathbf{w}_n), \mathbf{W}^{(2)}, \mathbf{W}^{(4)}, \mathbf{W}^{(l)}, \mathbf{W}^{(r)}, \mathbf{W}^{(sl)}, \mathbf{W}^{(sr)} \}, \quad (18)$$

具体来说,参数为嵌入词 $E \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$, 偏倚向量 $\mathbf{b}^{(2)} \in \mathbb{R}^H, \mathbf{b}^{(4)} \in \mathbb{R}^O$, 初始上下文 $\mathbf{c}_l(\mathbf{w}_1), \mathbf{c}_r(\mathbf{w}_n) \in \mathbb{R}^{|\mathcal{V}|}$, 变换矩阵 $\mathbf{W}^{(2)} \in \mathbb{R}^{H \times (|\mathcal{V}| + 2|\mathcal{V}|)}, \mathbf{W}^{(4)} \in \mathbb{R}^{O \times H}, \mathbf{W}^{(l)}, \mathbf{W}^{(r)} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}, \mathbf{W}^{(sl)}, \mathbf{W}^{(sr)} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$, 其中: $|\mathcal{V}|$ 是词汇表中的单词数; H 是隐藏层大小; O 是文档类型数。

采用随机梯度下降法对训练目标进行优化。在每个步骤中,随机选择一个例子 (D, class_D) 并进行梯度步骤

$$\theta \leftarrow \theta + \alpha = \frac{\partial \log p(\text{class}_D | (\mathbf{y}_k^{(4)}))}{\partial \theta}, \quad (19)$$

其中, α 是学习率。训练采用训练随机梯度下降的神经网络的方法。

在这部分,使用 skip-gram 模型预先训练嵌入的单词。最后通过最大化平均对数概率训练单词 $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_T$ 的嵌入

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(\mathbf{w}_{t+j} | \mathbf{w}_t), \quad (20)$$

$$p(\mathbf{w}_b | \mathbf{w}_a) = \frac{\exp(\mathbf{e}'(\mathbf{w}_b)^\top \mathbf{e}(\mathbf{w}_a))}{\sum_{k=1}^{|\mathcal{V}|} \exp(\mathbf{e}'(\mathbf{w}_k)^\top \mathbf{e}(\mathbf{w}_a))}, \quad (21)$$

其中: $|\mathcal{V}|$ 是未标记文本的词汇。 $\mathbf{e}'(\mathbf{w}_i)$ 是 \mathbf{w}_i 的一个嵌入。

Pooling层是特征提取层,将不同长度的文本转换为固定长度矢量,通过 pooling 层获取文本分类所需信息,研究使用最大池层,在文本中找到对分类准确率提升最有效率的词,池层利用循环结构的输出作为输入,最后利用 softmax 层进行分类。

3 实验结果及分析

本次实验使用 Windows10 编译 Python 程序,并在 GPU 上进行实验。选择的分类数据集为英文的 IMDB 电影评论的情感分析数据集,本数据集总共有 3 个部分组成,使用为带标签的训练集,共有数据 25 000 条,同时数据集中还包括 50 000 条不带标签的训练集以及 25 000 条数据的测试集。数据集中的 id 代表电影评论的 id, review 代表电影评论的内容, sentiment 代表情感分类的标签(只在带标签的数据集中存在)。

3.1 数据预处理

数据集为英文格式,不需要进行文本分词,只需要去掉停用词即可。停用词即出现频率比较高,但没有统计意义的词。通常将限定词视为停用词如:“the”、“a”、“that”、和“those”,这些词在文本中仅仅起到描述名词和表达概念的作用,对文本本身的内容并没有作用。介词如:“on”,“over”,“beneath”等表示相对位置,在文档处理中需要很大的空间,并且这些词的存在很普遍,只是丰富了文本,对文本分类的准确率没有较大提升。将停用词用列表的形式生成,方便之后查找停用词。

3.2 生成 word2vec 词向量

使用计算机进行文本分类,首先需要对使用的文字建模,适用于计算机处理。自然语言的建模方法从一般使用基于规则的方法到后期使用基于统计的方法,随着发展深度学习,建模方法也变得多种多样,特别是基于统计的方法受到广泛发展,如:n-gram、神经网络以及 log_linear 模型等建模方法。通常来说,建模的问题主要出现在建模维度过大、词的相似性难以区分、模型的针对性强等方面。为解决这些问题,模型的优化也在自然语言处理占了举足轻重的位置。对统计语言模型进行研究的背景下,Google 推出 Word2vec 训练词向量,能够通过给定的语料库,快速准确地将词转成向量的形式,对自然语言处理的发展打下了坚实基础。

word2vec 依托于浅层神经网络, 依赖 skip-grams 或 CBOW 建立神经词嵌入。词袋模型(Bag-of-words model) 是在自然语言处理和信息检索(IR)下被简化的表达模型。在文本处理中的作用主要是将文字整理成不考虑文法和顺序的词向量形式。CBoW 模型能将输入词的上下文变成输出这个特定词的词向量格式。由于词袋模型对于文本处理的优势, 在很多分类问题中都体现出较好性能, 词出现的频率可以用来当作训练分类器的特征。Skip-gram 模型是一个简单实用的模型。通常在解决自然语言处理的相关问题时, 语料选取特别重要。首先, 要解决语料不足的情况, 特别是需要反应词语关系的句子, 其次需要解决语料准确性问题, 输入与输出的对应关系对于文本问题的处理十分重要。CBoW 使用在小型数据库, Skip-Gram 使用在大型语料中。

这部分需要生成词向量和词汇-索引映射字典, 去掉停用词, 进行词频的统计, 将统计出来的低频词删除, 最后构建适于文本的专属词典, 将处理后的数据与词典带入 Word2vec 模型, 构建词向量。

3.3 基于改进 LSTM 模型的文本分类

利用预训练的词向量初始化词嵌入矩阵, 利用词嵌入矩阵将输入数据中的词转换成词向量, 复制一份词嵌入层输入, 嵌入层可以对词向量处理, 使得维度降低, 减少运算量。将处理后的数据输入 Bi-LSTM 结构, 将前向, 后向的输出和最早的词向量拼接在一起得到最终词表征, 做 max-pooling 的操作, 将时间步的维度消失, 最后全连接层输出分类结果。

模型与循环神经网络模型, 长短时记忆网络模型, 带注意力的长短时记忆网络模型进行对比实验。为充分展示改进模型分类优势, 实验中各模型均采用表 1 给出的 3 组参数。参数的设置由经验选择了以下 4 个参数进行实验。

表 1 实验中采用的 3 组参数设置

Tab.1 Three sets of parameter settings used in the experiment

参数名称	设置 1	设置 2	设置 3
迭代次数	1 000	2 000	1 000
隐层神经元个数	128	128	128
学习率	0.001	0.001	0.003
嵌入层神经元个数	200	200	200

对于每组参数设置, 各模型分别独立运行 50 次, 对测试集的平均识别率, 随迭代步数动态变化的曲线如图 3~5 所示, 然后将分类正确率的平均值作为对比指标, 对比结果如表 2 所示。

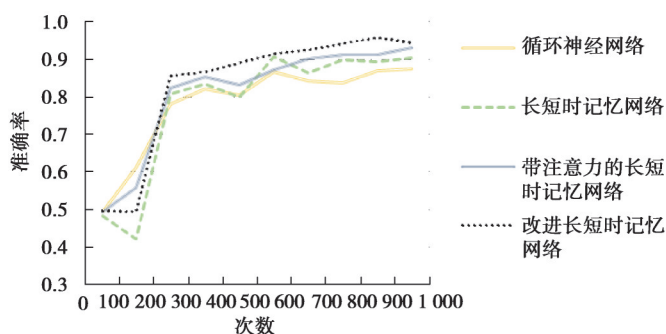


图 3 设置 1 各模型的准确率与迭代次数关系

Fig.3 The relationship between the accuracy of each model and the number of iterations of set 1

3.4 实验结果分析

图 3 的结果显示在迭代次数 0~350 代时, 改进的长短时记忆网络在准确率上并没有很好的结果, 但随着迭代次数加大, 准确率比其他模型优秀, 体现改进 LSTM 需要更长期的迭代才能体现优势, 特别在图 4 中发现, 在 1 000 次迭代后改进的 LSTM 优势变小, 说明随着代数增加到一定程度, 循环神经网络的限制显示出

来,而在加大了学习率的设置中,循环神经网络很快到达准确率较高的区域,也很快趋于平缓,相对来说,改进的 LSTM 在学习率变化时准确率的变化较稳定。通过表 2 的结果对比可以发现,在相同的参数影响下,改进长短时记忆网络对文本分类的准确率均高于传统的长短时记忆网络,证明对传统单向网络的双向改造是成功的,双向长短时记忆网络不仅可以学习它的正向规律,还可以学习它的反向规律,通过将正反向结合进行分类,发现分类效果更佳优秀,对于文本分类问题来说,充分理解整句话,对于分类的准确度具有十分重要作用,对整句话进行处理后,最大池化层可以帮助选取对分类更重要部分,找到最重要潜在语义因素,从而提高分类的准确性。

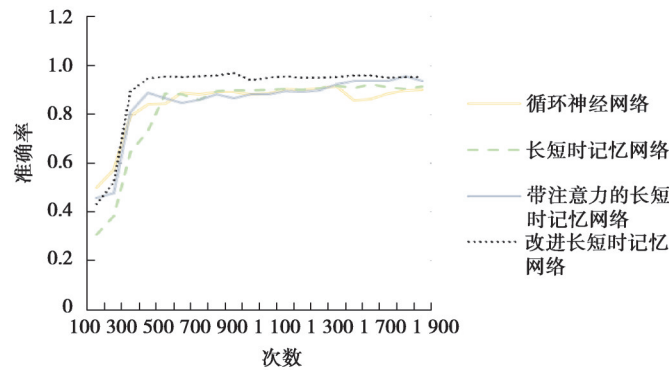


图 4 设置 2 各模型的准确率与迭代次数关系

Fig.4 The relationship between the accuracy of each model and the number of iterations of set 2

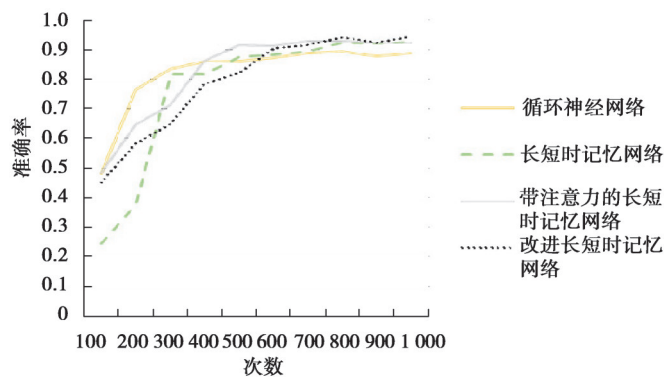


图 5 设置 3 各模型的准确率与迭代次数关系

Fig.5 The relationship between the accuracy of each model and the number of iterations of set 3

表 2 50 次实验的平均结果对比

Tab.2 Comparison of average results of 50 experiments

模型	IMDB 评论数据		
	设置 1	设置 2	设置 3
循环神经网络	87.4	90.2	88.9
LSTM	90.4	91.5	92.7
带注意力的 LSTM	93.2	93.7	92.4
改进 LSTM	94.6	95.3	94.5

%

接下来,将会考虑在引入对抗学习进行文本分类^[5]和引用弱监督和深度表示的文本分类^[6]方面进行改进。考虑到所用训练数据为 25 000 条标记的数据,下一步拟通过将部分未标记的数据加入到训练集中,进而采用半监督学习的方式来进一步提升文本分类的准确率;同时,拟采用迁移学习研究该模型面向中文的文本分类。

4 结 论

研究通过2次改进,提高文本分类的效果。一方面将传统的长短时记忆网络拓展为可双向捕捉文本重点的双向长短时记忆网络,根据文本的整体含义来进行分类,增强对数据的利用效果;另一方面在分类前增加了最大池化层,使得对文本分类更重要的词能够发挥其对分类的影响。综合以上,研究提出的模型成功提高了分类的成功率。

参考文献

- [1] 石锋. 面向中文新闻文本的实体关系抽取研究[D]. 哈尔滨: 哈尔滨工业大学, 2017.
Shi F. Research on entity relationship extraction for Chinese news text[D]. Harbin: Harbin Institute of Technology, 2017. (in Chinese)
- [2] 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.
- [3] 崔莹. 深度学习在文本表示及分类中的应用研究[J]. 电脑知识与技术, 2019, 15(16): 174-177.
Cui Y. Application of deep learning in text representation and classification with deep learning[J]. Computer Knowledge and Technology, 2019, 15(16): 174-177.(in Chinese)
- [4] Liu Q, Yuan J Z, Weng C H. Survey of short text classification based on deep learning[J]. Computer Sciences, 2017, 44(A), 11-15
- [5] Bi H, Sun J, Xu Z. A Graph-Based semisupervised deep learning model for polSAR image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, PP(99):1-17.
- [6] 陈文实, 刘心惠, 鲁明羽. 面向多标签文本分类的深度主题特征提取[J]. 模式识别与人工智能, 2019, 32(9): 785-792.
Chen W S, Liu X H, Lu M Y. Feature extraction of deep topic model for multi-label text classification[J]. Pattern Recognition and Artificial Intelligence, 2019, 32(9): 785-792.(in Chinese)
- [7] Lee J, Yu I, Park J, et al. Memetic feature selection for multilabel text categorization using label frequency difference[J]. Information Sciences, 2019, 485: 263-280.
- [8] 牛硕硕, 柴小丽, 李德启, 等. 一种基于神经网络与LDA的文本分类算法[J]. 计算机工程, 2019, 45(10): 208-214.
Niu S S, Chai X L, Li D Q, et al. A text classification algorithm based on neural network and LDA[J]. Computer Engineering, 2019, 45(10): 208-214.(in Chinese)
- [9] 向进勇, 杨文忠, 吾守尔·斯拉木. 基于特征选择和深度信念网络的文本情感分类算法[J]. 计算机应用, 2019, 39(7): 1942-1947.
Xiang J Y, Yang W Z, Silamu W. Text sentiment classification algorithm based on feature selection and deep belief network[J]. Journal of Computer Applications, 2019, 39(7): 1942-1947.(in Chinese)
- [10] 马力, 李沙沙. 基于词向量的文本分类研究[J]. 计算机与数字工程, 2019, 47(2): 281-284, 303.
Ma L, Li S S. Research on text classification based on word embedding[J]. Computer & Digital Engineering, 2019, 47(2): 281-284, 303.(in Chinese)
- [11] 陈巧红, 王磊, 孙麒, 等. 卷积神经网络的短文本分类方法[J]. 计算机系统应用, 2019, 28(5): 137-142.
Chen Q H, Wang L, Sun Q, et al. Short text classification based on convolutional neural network[J]. Computer Systems & Applications, 2019, 28(5): 137-142.(in Chinese)
- [12] 涂文博, 袁贞明, 俞凯. 针对文本分类的神经网络模型[J]. 计算机系统应用, 2019, 28(7): 145-150.
Tu W B, Yuan Z M, Yu K. Neural network models for text classification[J]. Computer Systems & Applications, 2019, 28(7): 145-150.(in Chinese)
- [13] Deng H L, Zhang L, Shu X. Feature memory-based deep recurrent neural network for language modeling[J]. Applied Soft Computing, 2018, 68: 432-446.
- [14] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
- [15] Poon H K, Yap W S, Tee Y K, et al. Hierarchical gated recurrent neural network with adversarial and virtual adversarial training on text classification[J]. Neural Networks, 2019, 119: 299-312.
- [16] Wang Y S, Sohn S, Liu S J, et al. A clinical text classification paradigm using weak supervision and deep representation[J]. BMC Medical Informatics and Decision Making, 2019, 19(1): 1.

(编辑 侯 湘)