基于BGRU池的卷积神经网络文本分类模型

周 枫 李荣雨
(南京工业大学计算机科学与技术学院 南京 211816)

摘 要 针对深度学习在处理文本分类问题时存在的适应度小、精确度较低等问题,提出一种采用双向门控循环单元(BGRU)进行池化的改进卷积神经网络模型。在池化阶段,将BGRU产生的中间句子表示与由卷积层得到的局部表示进行对比,将相似度高的判定为重要信息,并通过增大其权重来保留此信息。该模型可以进行端到端的训练,对多种类型的文本进行训练,适应性较强。实验结果表明,相较于其他同类模型,提出的改进模型在学习能力上有较大优势,分类精度也有显著提高。

关键词 深度学习,卷积神经网络,双向门控循环单元,文本分类
中图法分类号 TP183,TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.06.042

Convolutional Neural Network Model for Text Classification Based on BGRU Pooling

ZHOU Feng LI Rong-yu
(School of Computer Science and Technology, Nanjing Tech University, Nanjing 211816, China)

Abstract Aiming at the problem that deep learning has the disadvantages of small adaptability and low precision when it solves the problem of text classification, this paper proposed a convolution neural network model based on bi-directional gated recurrent unit (BGRU) and convolution layer pooling. In the pooling stage, the intermediate sentence generated by BGRU is represented as a local representation obtained from the convolution layer, the representation of high similarity is judged to be important information, and the information is retained by increasing its weight. The model can give end-to-end training and train multiple types of text, and it has good adaptability. The experimental results show that the proposed model has greate advantage compared with other similar models, and the classification accuracy is also improved significantly.

Keywords Deep learning, Convolutional neural network, Bi-directional gated recurrent unit, Text classification

1 引言

文本分类问题是自然语言处理(Natural Language Processing,NLP)领域内一个非常经典的问题,可被应用于情感分析、文献综述、机器翻译、语篇分析等任务中。文本分类问题的相关研究最早可以追溯至20世纪50年代,当时是通过专家规则(Pattern)[1]进行分类,不仅费时费力,且准确率也十分有限,完全不具备推广性。另外,随着互联网的发展,在线文本数量急剧增长,传统方法难以处理如此庞大的数据量。新兴的深度学习因能处理海量的数据且易于实现[2-3],成为了解决文本分类问题的一种重要手段。

深度学习在解决文本分类任务时可分为特征工程和分类器两部分。特征工程又分为文本预处理、文本表示、特征提取3个部分,其最终目标是把文本转换成计算机可理解的格式,并封装足够的用于分类的信息,因此具有很强的特征表达能力。分类器大多选用统计学方法,但选用何种算法根据具体情况而定,本文因着眼于特征工程部分,对其不予讨论。处理特征工程最经典的方法是词袋模型,但其由于失去了词序这一重要因素,不能处理语义分析的问题,且存在高维度和高稀疏性两个问题,近年来逐渐被分布式表示所取代。目前最常用的分布式表示为Word Embeddings[16],它具有低维度、连续向量、易处理且可以在向量空间中捕捉语义和句法信息等优点。

经由Word Embeddings处理过的文本,虽具有一定的特征表达能力,但还不足以应对实际运用。在此基础上,学者们提出利用神经网络模型来建立非线性词语之间的相互作用关系,通过学习句子的表述来捕捉句子的语义。目前,循环神经网络逐渐成为文本分析领域的一大热点,其表现形式为网络会对前面的信息进行记忆,并将其用于当前输出的计算中。理论上,循环神经网络能够对任何长度的序列数据进行处理,但该网络面临着梯度消失或梯度激增的问题,容易降低基于梯度的优化方法的效率,甚至使其无效。为了解决这个问题,Kim等人提出了一种卷积神经网络模型[4],其通过对超参数进行调整并采用静态载体,在情感分析和问题分类任务中表现优异,但它只能在特定任务和静态向量的体系结构中使用,适用性不强。Kalchbrenner等人对池化层进行修改[5],通过

到稿日期:2017-05-03 返修日期:2017-09-08 本文受江苏省高校自然科学基金资助项目(12KJB510007)资助。
周 枫(1992—),男,硕士生,主要研究领域为机器学习与深度学习;李荣雨(1977—),男,博士,副教授,主要研究领域为面向流程工业的机器学习,E-mail:alwayslry@sina.com(通信作者)。

将池化层改成 k-max pooling，即可在池化阶段保留 k 个最大的信息，由此保留了序列信息。然而，这个网络只能保留局部信息，没能充分利用上下文。Qin 等人[6]提出了一种可以从数据本身自动学习特征的卷积网络结构模型，该网络可处理不同长度的输入句子，而且能够捕捉短期和长期的语义关系，在多类情感预测问题中表现出色；但由于是无监督网络，其精确度较低。

本文通过在池化层融入双向门控循环单元(Bi-directional Gated Recurrent Unit，BGRU)，提出了一种基于 BGRU 的改进卷积神经网络模型。该模型能够提高卷积层提取综合信息的能力，在池化阶段保留最显著的信息，能有效地解决现有卷积神经网络池化层因易丢失特征的位置信息和强度信息而导致学习性能降低的问题。改进模型可以更多地保留句子中包含的信息，产生一个更具代表性的特征向量，从而得到更好的分类结果。

2 相关介绍

2.1 卷积神经网络

卷积神经网络(Convolutional Neural Network，CNN)是一种深度学习网络，最早由 Lecun 于 1998 年提出[7]，通常由特征提取层、下采样层以及一个顶层分类器组成。该网络最初被应用于手写字体识别任务中，随后在图像识别、语言识别等领域得到了广泛应用。CNN 在图像上取得优异性能的重要原因之一在于其特征提取能力，即当多层卷积层堆叠时，其可以从原始图像像素这样的低级特征中逐渐提取到边缘、角点、轮廓等，直到整个识别目标。而这种特征的层级表示现象不只在图像数据中存在，文本中从字到词、短语、句子、段落等逐层递进过程也体现了特征的层级表示现象，因此 CNN 也可以用于文本数据[17]。

2.2 池化层

池化层是 CNN 的一个重要组成部分，其输入一般来源于上一个卷积层，可提供很强的鲁棒性，并且能减少参数的数量，防止过拟合。池化层一般没有参数，因此反向传播时只需对输入参数求导，不需要进行权值更新。1-max pooling 被应用于整个特征映射图中，通过池化操作，每个滤波器固定取 1 个值，即有多少个滤波器，池化层就有多少个神经元，这样就可以把全联接层的神经元个数固定住。但是，1-max 也存在缺点：首先，特征的位置信息在这一步骤完全丢失，虽然在卷积层保留了特征的位置信息，但是通过取唯一的最大值，池化层便只知道这个最大值是多少，而丢失了位置信息；其次，若某些强特征出现多次，但是 1-max 只保留一个最大值，则即使某个特征出现多次，也只能提取出一次，即该特征的强度信息丢失了。Kalchbrenner 等人提出了一种有效的共享策略——k-max pooling[5]，它可以取特征值中所有得分在 Top-k 的值，并保留这些特征值的原始先后顺序，即多保留一些特征信息以供后续阶段使用。可以看出，1-max pooling 是 k-max pooling 的一种特殊情况。除此之外，还有 average pooling(返回特征映射的平均值)以及 local max pooling(连接每个局部区域生成的数字而形成特征向量)等[9-10]。

2.3 BGRU

区别于其他神经网络，循环神经网络(Recurrent Neural Networks，RNN)中隐藏层之间的节点是连接的，且隐藏层的输入同时包含了来自输入层的输出以及上一时刻隐藏层的输出，这种特性使得 RNN 具有一定的记忆功能。理论上，RNN 可对任何长度的序列数据进行处理；但在实际运行过程中，如果间隔时间过长，RNN 将失去之前的信息，且其复杂性也会上升。对此，学者们提出了几种改进方案。

长短时记忆(Long-short Memory，LSTM)和门控循环单元(Gated Recurrent Unit，GRU)是其中较为常见且效果较好的两种改进方案。LSTM 由 Hochreiter 等人[11]提出，由输入门、遗忘门、输出门以及一个记忆核心(cell)组成。GRU 由 Chung 等人于 2014 年提出[12]，将 LSTM 中的遗忘门与输入门合成一个更新门，同时没有单独的存储单元。与 LSTM 相比，GRU 的结构更简单，对于相同的任务所需的收敛时间更短，但是训练效果却与 LSTM 相差无几。因此，本文选用 GRU，它可以在每个循环单元中自适应地捕获不同时间尺度的依赖关系。

GRU 在 t 时刻的激活值 h_t^j 是 $t-1$ 时刻的激活值 h_{t-1}^j 和候选激活值之间的线性插值，计算公式如下所示：

$$h_t^j=(1-z_t^j)h_{t-1}^j+z_t^j\,\tilde{h}_t^j \tag{1}$$

其中，z_t^j 为更新门，决定在此次运算中是否忽略当前词 x_t。更新门由式(2)得出：

$$z_t^j=\sigma(W_z x_t+U_z h_{t-1})^j \tag{2}$$

其中，W_z 是第 t 次输入的词向量 x_t 对应的系数矩阵，U_z 是 $t-1$次输出 h_{t-1}对应的系数矩阵，σ 是激活函数 sigmoid。候选激活值的计算公式如式(3)所示：

$$\tilde{h}_t^j=\tanh(Wx_t+U(r_t\odot h_{t-1}))^j \tag{3}$$

其中，r_t 是重置门，$\odot$是点乘运算。从式(3)中可以看出，当 r_t 接近于 0 时，候选激活值会忽略之前隐藏的节点信息，即可忘记之前的计算状态。此机制可使模型抛弃某些无用信息，降低计算复杂度。重置门 r_t 的计算公式如下：

$$r_t^j=\sigma(W_r x_t+U_r h_{t-1}) \tag{4}$$

GRU 模型的门控机制使记忆单元在工作中可以保持一段时间的信息记忆，并避免在训练时其内部梯度受到不利因素的干扰。因此，GRU 适合处理和预测时间序列中间隔和延迟非常长的重要事件。鉴于 GRU 只能获取单向文本信息，本文设计了一种 BGRU 模型，以期从两个方向来获取上下文信息。具体模型如图 1 所示，其中 G 为 GRU 模块。

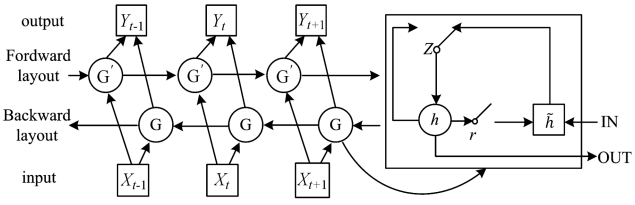

图 1 双向门控循环单元

Fig. 1 Bi-directional gated recurrent unit

3 基于 BGRU 的 CNN 模型

本节详细描述了本文所提出的模型。图 2 给出改进模型的结构图。将通过 Word Embeddings 生成的词向量同时送入 BGRU 与卷积层，卷积滤波器对输入的词向量进行卷积，

以生成局部表示,卷积运算可以捕捉句子中每个位置所包含的局部信息。BGRU 池层被用来整合局部特征,以组成包含权重的文本表示。该权重通过将局部表示与 BGRU 在训练阶段生成并优化的中间句子表示进行比较而得出。然后,将所有卷积滤波器生成的文本表示连接成最终的句子表示,并将其送入顶层的 Softmax 分类器。由 BGRU 产生的中间句子特征也将在测试阶段作为 Softmax 分类器的输入,如图 2 中虚线所示。

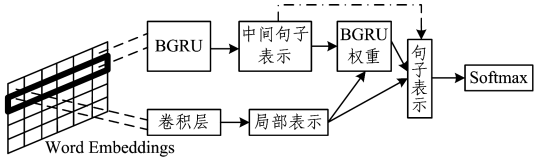

图 2　基于 BGRU 的卷积神经网络模型

Fig. 2　Convolution neural network model based on BGRU

3.1　Word Embeddings

引入词向量是为了将文本转换为计算机能够识别的形式,从而可以使各种算法应用于 NLP 任务。在此之前,最经典的词向量表达方式为 one-hot 编码,其在文档分类任务中具有良好的学习表现。然而,该编码方式在分类短文本时会使得生成的词向量非常稀疏,难以保留词与词之间的语义关系,有可能会面临维度灾难问题。为此,Hinton 提出了一种新的词向量表达方式——Word Embeddings,其原理为将词分布式地映射到低维空间中,不仅可以很好地解决词向量稀疏问题,也可通过该空间中词与词之间的位置关系反映它们在语义上的联系。

为了更好地利用语义和语法信息,本文采用 Word2vec 嵌入所有数据集。通过 skip-gram 方法训练数据,如式(5)和式(6)所示:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leqslant j\leqslant c,j\neq 0}\log p(x_{t+j}\mid x_t) \tag{5}$$

$$p(y\mid x)=\frac{\exp(x_i^{\mathrm{T}}y)}{\sum_{j=1}^{V}\exp(x_j^{\mathrm{T}}y)} \tag{6}$$

其中,c 代表窗函数的大小;$x_1,x_2,\cdots,x_T$ 代表某一词组序列。

3.2　卷积层

本文模型中的卷积运算通过 K 型滤波器 $W_c\in R^{md*k}$ 和级联向量 $x_{i:i+m-1}$(一个窗口包含从第 i 个词开始的 m 个词)在相应的特征图中获得该窗口中词的特征。d 是 Word Embeddings 的维数,每个过滤器的参数在所有窗口共享。使用具有不同初始化权重的多个过滤器来改进模型的学习能力,使用交叉验证来确定滤波器的数量 k。卷积运算如式(7)所示:

$$c_i=g(W_c^{\mathrm{T}}x_{i:i+m-1}+b_c)\in R^k \tag{7}$$

其中,$x_i\in R^d$,b 和 c 是偏差向量,$g()$是非线性激活函数。本文采用 ReLU 作为激活函数,因为其可以有效改善网络的学习效率并显著减少深度学习网络中收敛所需的迭代次数,近来已经成为 CNN 的标准非线性激活函数。

假设句子的长度为 T,本文将卷积的边界模式设置为"相同",即使得卷积层的输出长度与输入长度相同。如果边界长度不足,那么通过零填充来保证相同的长度。当词窗口滑动时,卷积层的特征图表示如式(8)所示:

$$c=[c_1,c_2,\cdots,c_T]\in R^{k*T} \tag{8}$$

卷积层的输出为句子的局部表示,每个元素 c_i 是对应位置的局部表示。

3.3　BGRU-CNN

现有的卷积神经网络大多使用 max-pooling 作为池化层,但是 max-pooling 丢失了特征的位置信息和强度信息,会降低学习性能。针对这一问题,本文提出一种新的池策略。如图 1 所示,通过 BGRU 生成一个中间句子表示。BGRU 是循环神经网络的变体,它可以学习序列中包含的历史信息和未来信息,还能通过用门控存储器单元替代回归神经网络的隐藏状态来解决"梯度消失"的问题。本文将中间句子表示设为 $\tilde{s}$。

一旦生成了中间句子表示,就可以通过与卷积层生成的局部表示进行对比来计算权重。为了比较两个表示,本文通过控制 BGRU 的输出尺寸使之与 K 型卷积滤波器的数目相同,将局部表示和中间句子表示映射到相同维度的空间。中间句子表征和每个局部表征之间的相似性越高,则该局部表示就应当被分配更大的权重。该权重的计算公式如式(9)、式(10)所示:

$$\partial_i=\frac{\exp(e_i)}{\sum_{i=1}^{T}\exp(e_i)} \tag{9}$$

$$e^i=sim(c_i,\tilde{s}) \tag{10}$$

其中,∂_i 是一个标量,函数 $sim()$通过余弦函数来测量两个输入之间的相似度。在得到权重后,句子表示 s 由式(11)得到:

$$s=\sum_{i=1}^{T}\partial_i c_i\in R^k \tag{11}$$

该 BGRU 模型参与了模型中所有组件的训练。代价函数的反向传播梯度通过中间句子表示,以便在训练阶段对其进行优化。因此,本文模型可以实现端到端的训练。由 BGRU 模型得到的中间句子表示将与卷积层得到的句子表示作为测试阶段顶部分类器的输入。

BGRU 池可以看作是所有单词注释的加权和,可以用来诠释整个句子。每个单词的权值大小可用于衡量单词对整个句子的意义大小。这种方法借鉴了最近在机器翻译、物体识别等领域有出色表现的"专注"概念[13],即该模型可以决定哪些特征需要增大权值。与最大池法相比,BGRU 池法能够保留更多句子包含的信息。与平均池法相比,新的池策略可以将更大的权重分配给更重要的特征,这显然更具优势。

BGRU 和卷积结构组合是本文模型的重要组成部分。由 BGRU 产生的中间句子表示,已经足以作为顶层分类器的输入表征。而本文通过比较局部表征与中间句子表征,所获得的权重编码能保留更多的句子信息。另一方面,与只使用 BGRU 相比,加入卷积结构,并充分利用它的局部上下文提取能力,可以有效地提高信息检索能力。这使得该模型可以访问更全面的信息,即包含历史、未来和局部上下文中任何位置的信息。简而言之,位置和情感极性信息在新机制下被完全保留了下来。

顶部分类器的输入在训练阶段是句子表示 s,而在测试阶段则使用$[s,\tilde{s}]$。通过线性变换层和 Softmax 层在模型顶

层相加，得到类空间上的条件概率。为了避免过拟合，在倒数第二层加入有屏蔽概率 p 的 dropout。dropout 的关键思想是在训练阶段从神经网络中按概率 p 随机丢弃单元，被丢弃的神经元不参与此次训练。因此，输出层的计算如式(12)、式(13)所示：

$$y=\begin{cases}W_s(s\odot q)+b_s, & \text{训练阶段}\\ W_s([s,\tilde{s}])+b_s, & \text{测试阶段}\end{cases} \tag{12}$$

$$P_c=\frac{\exp(y_c)}{\sum_{c'\in C}\exp(y_{c'})} \tag{13}$$

其中，$\odot$ 是点乘运算，q 是带有随机丢弃单元概率 p 的屏蔽向量，c 是类的数量。另外，通过 L2 范数来限制输出权重 w_s。本文采用的是有监督模型，故每个句子都有标签 P_c^g。式(14)是根据最小化分类交叉熵得到的目标函数：

$$L=-\sum_{i=1}^{N}\sum_{c=1}^{C}P_c^g(S_i)\log(P_c(S_i)) \tag{14}$$

其中，P_c^g 采用 one-hot 编码方案。由此模型确定的参数包括卷积滤波器、BGRU 和 Softmax 分类器中的所有权重和偏置项，BGRU 权重将在训练阶段被更新。该模型的算法描述如算法 1 所示。

算法 1 BGRU-CNN

1. 该算法的输入为 N 个可变长度的句子及其相应的标签 P_c^g，标签已由 one-hot 编码方案表示；
2. 对某个具体的句子而言，使用式(5)预训练出的词向量来构造句子矩阵；
3. for i in [1,H] do
4. 窗口大小为 m^i 的第 i 个卷积神经网络，根据式(8)，利用一个卷积层来获得局部表示 $c=[c_1,c_2,\cdots,c_T]$；
5. 使用 BGRU 来获得中间句子表示；
6. 使用式(9)和式(10)来计算 BGRU 权重；
7. 通过式(11)将 BGRU 权重与局部表示相结合，以获得句子表示；
8. end for
9. 将 H 句表征连接起来，形成最终的句子表示；
10. 将最终的句子表示输入到 Softmax 分类器中，以预测类标签；
11. 使用所有训练好的句子和标签，通过式(14)来更新模型的参数；
12. 在测试阶段，将中间句子表示、由卷积层得到的局部表示以及 BGRU 权重三者结合，通过 Softmax 分类器输出最终的分类结果。

4 实验结果和分析

本节评估了所提出的模型在 6 个基准数据集上的文本分类的性能，并将其与其他方法进行比较。

4.1 基准数据集

1)MR：电影评论的情感极性数据集，用于评估影评文本整体情感极性的正负性。正、负数据量分别为 5331，5331。

2)SUBJ：主观性数据集，用于分析某个句子的主客观性。主、客观数据量分别为 5000，5000。

3)CR：顾客评论数据集，用于将每个评论分类为正面的或负面的。数据分布为 2411，1373。

4)MPQA：包含 535 篇各个来源的文章，在句子层和子句层已经手工标注好了观点和其他属性(如信仰、情感、揣测等)，本文筛选出情感极性部分，数据分布为 3310，7296。

5)SST-1：斯坦福大学情感数据库，是 MR 数据库的一个延伸。此数据库提供了从训练到开发再到测试的过程中所需的标签(非常积极的、积极的、中性的、消极的、非常消极的)。数据分布为 1837，3118，2237，3147，1516。

6)SST-2：由 SST-1 延伸而来，只包含两个标签，即积极的和消极的，不包含中性的数据。数据分布为 4955，4663。

4.2 参数设置

本文使用 word2vec 训练词向量，每个词向量的维数为 300。在训练阶段会对词向量进行一定的微调，以使其适应每个任务。实验过程中参数的设置会对网络性能产生很大的影响[14]，诸如滤波器窗口大小和特征图的数量。为此，对每个参数进行单独测试。为了方便对比，以窗函数大小为 3，dropout 率为 0.5，L2 范数约束条件为 3，特征图谱数为 100 作为基准设置，即当测试某一参数时，其余参数如上所示，测试结果如图 3 所示。

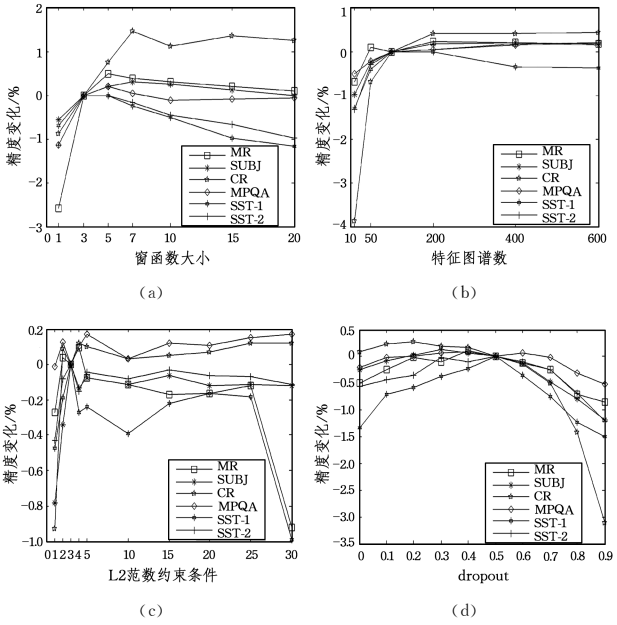

图 3 参数配置对模型精确度的影响

Fig. 3 Effect of parameter configuration on model accuracy

测试基准数据集中 4 个参数的设置如表 1 所列。实验过程中发现，当这 4 个参数处于适当的范围时，其数值大小不会影响最终性能。每个数据集的 BGRU 的输出尺寸设置与特征图的数量相同，以便将局部表示与中间句子表示进行比较。因为 SST-1 和 SST-2 的训练数据量非常大，所以将 SST-1 和 SST-2 的训练批量的大小设置为 100，而其他数据集的训练批量设置为 50。学习率、衰减因子和 Adadelta 模糊因子的设置分别为 1.0，0.95 和 0.000001。

表 1 基准数据集

Table 1 Benchmark dataset

数据集	窗口大小	特征图谱	Dropout	L2 范数
MR	5	200	0.5	4
SUBJ	7	200	0.4	3
CR	7	200	0.3	4
MPQA	5	100	0.4	5
SST-1	3	100	0.5	3
SST-2	3	100	0.5	3

4.3　对比实验

4.3.1　与其他算法进行对比

将 BGRU-CNN 模型与其他算法进行对比，结果如表 2 所列。从原始文献中提取 NB-SVM，MNB，RAE，MV-RNN，RNTN，CNN，one-hot CNN 和 DCNN 的结果[15]；而 CBOW，RNN 和 BRNN 数据则由仿真得到，并从中选用最佳结果。从表中可以看出，短句的 n-gram 编码的稀疏性使得两个贝叶斯模型在有较长句子的数据集上表现优异，但在短句数据集上的表现不佳。CBOW 由于在句子中丢失了单词顺序信息，其精确度也相对较低。3 个递归神经网络结构，即 RAE，MV-RNN 和 RNTN，由于其表现在很大程度上依赖于解析树的建设，且受过拟合问题的影响，因此分类准确性较低，并不能得到令人满意的性能。另外，从表 2 中可以看出，与 RNN 相比，BRNN 的学习效果更好，从而证明了双向结构有助于提取更多信息，能增强学习效果。

表 2　算法精确度的对比

Table 2　Comparison of algorithmic accuracy

（单位：%）

算法	MR	SUBJ	CR	MPQA	SST-1	SST-2
NB-SVM	79.4	93.2	81.8	86.3	—	—
MNB	79.0	93.6	80.0	86.3	—	—
CBOW	77.2	91.3	79.9	86.4	42.8	81.5
RAE	77.7	—	—	86.4	43.2	82.4
MV-RNN	79.0	—	—	—	44.4	82.9
RNTN	—	—	—	—	45.7	85.4
RNN	77.2	92.7	82.3	90.1	47.2	85.8
BRNN	81.6	92.3	82.6	90.3	48.1	86.5
CNN	81.5	93.4	84.3	89.5	48.0	87.2
one-hot CNN	77.8	91.1	78.2	83.9	42.0	79.8
DCNN	—	—	—	—	48.5	86.8
BGRU-CNN	82.3	94.4	86.0	90.5	50.2	89.5

从表 2 中可以得出，BGRU-CNN 在所有任务中都优于其他系统，从而证明这种策略是十分有效的，即 BGRU 模型和卷积结构相结合使得模型能够提取综合信息，包括句子序列中历史、未来和局部上下文的信息。新模型的有效性已通过实验得到验证。

4.3.2　池策略比较

本节通过与现有的池策略进行对比来验证基于 BGRU 的池策略的有效性。为了公平起见，将 local max pooling 的大小设置为 10，将 k-max pooling 的 k 值设置为 3，其余所有参数设置相同。精度对比如图 4 所示。

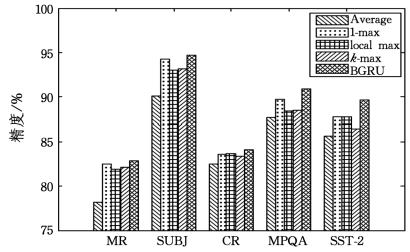

图 4　池策略精确度的对比

Fig. 4　Comparison of pooling policy precision

由于 SST-1 的精度范围远小于其他数据集，且 SST-1 的比较结果与 SST-2 相似，因此为了进行更好的比较，本次不选用 SST-1 数据集。从图 4 中可以看出，本文所提池策略在所有数据集中的表现均优于其他池策略。其中，average pooling 策略的效果最差，表明对所有特征设置相等的权重是不合理的。此外在大多数情况下，1-max pooling 策略的表现都比 k-max 和 local max pooling 好，尽管后者保留了更多的信息，所以在某些情况下，并不是保留越多的信息越好；而本文所提方案能提取句子中包含的最重要的信息，因此该方案是有效的。

结束语　本文针对文本分类任务提出的基于双向门控循环单元的卷积神经网模型是十分有效的。通过 BGRU 产生一个中间句子表示，可以将更重要的信息提取出来；且 BGRU 与卷积结构相结合，可以使模型能够提取一个句子中全面和最重要的信息。实验结果表明，本文模型在文本分类任务中的表现十分优异，可以实现端到端的训练，且能在不同的语义空间中对文本分类，适应性较好。

参 考 文 献

[1] SEBASTIANI F. Machine learning in automated text categorization[J]. Acm Computing Surveys, 2001, 34(1): 1-47.

[2] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[3] GUO L L, DING S F. Research Progress on Deep Learning[J]. Computer Science, 2015, 42(5): 28-33. (in Chinese)
郭丽丽，丁世飞. 深度学习研究进展[J]. 计算机科学，2015，42(5)：28-33.

[4] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv: 1408.5882, 2014.

[5] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P. A convolutional neural network for modelling sentences[J]. arXiv preprint arXiv: 1404.2188, 2014.

[6] QIN P, XU W, GUO J. An empirical convolutional neural network approach for semantic relation classification[J]. Neurocomputing, 2016, 190: 1-9.

[7] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[8] WANG P, XU B, XU J, et al. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification[J]. Neurocomputing, 2016, 174: 806-814.

[9] BOUREAU Y L, ROUX N L, BACH F, et al. Ask the locals: multi-way local pooling for image recognition[C]// 2011 IEEE International Conference on Computer Vision (ICCV). IEEE, 2011: 2651-2658.

[10] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[11] HOCHREITER S, SCHMIDHUBER J, et al. Long short-term memory[J]. Neural Computer, 1998, 9(8): 1735-1780.

[12] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [J]. arXiv preprint arXiv: 1412.3555, 2014.

[13] YONG Z,MENG J E,NING W,et al. Attention Pooling-based Convolutional Neural Network for Sentence Modelling[J]. Information Sciences,2016,373(C):388-403.

[14] GREFF K,SRIVASTAVA R K,KOUTNÍK J,et al. LSTM:A search space odyssey[J]. IEEE Transactions on Neural Networks and Learning Systems,2017,28(10):2222-2232.

[15] ZHANG Y,WALLACE B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1510.03820,2015.

[16] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781,2013.

[17] ZHANG C Y,QIN P D,YI Y L. Self-adaptation Multi-gram weight learning strategy for sentence representation based on convolutional neural network[J]. Computer Science,2017,44(1):60-64. (in Chinese)
张春云,秦鹏达,尹义龙. 基于卷积神经网络的自适应权重multi-gram语句建模系统[J]. 计算机科学,2017,44(1):60-64.

(上接第 210 页)

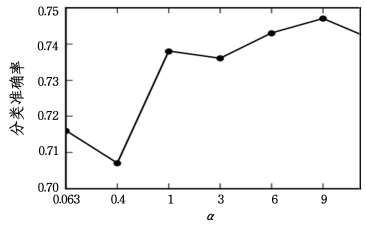

图 7 IAE 算法的分类准确率与 α 之间的关系

Fig. 7 Relationship between classification accuracy and α of IAE algorithm

结束语 词嵌入是文本挖掘中各个应用的基本前提,其效果严重影响着文本挖掘的准确性。本文对自编码器进行了改进,在隐藏层中加入了全局调整函数,实现了嵌入式特征向量的稀疏化,解决了文本数据的稀疏性问题,从而提高了其在后续分类应用中的准确性。最后,通过在 20news 数据集中的实验验证了所提算法的效果。

参 考 文 献

[1] ELLISON N B. Social network sites:definition,history,and scholarship[J]. Journal of Computer-Mediated Communication,2007,13(1):210-230.

[2] HOFMANN T. Probabilistic latent semantic analysis[C]// Fifteenth Conference on Uncertainty in Artificial Intelligence. 1999:289-296.

[3] SONG Y,PAN S,LIU S,et al. Topic and keyword re-ranking for LDA-based topic modeling[C]// 18th ACM Conference on Information and Knowledge Management. 2009:1757-1760.

[4] BROWN P F,DESOUZA P V,MERCER R L,et al. Class-based n-gram models of natural language[J]. Computational linguistics,1992,18(4):467-479.

[5] MIKOLOV T,SUTSKEVER I,CHEN K,et al. Distributed representations of words and phrases and their compositionality[C]// 26th International Conference on Neural Information Processing Systems. 2013:3111-3119.

[6] LE Q,MIKOLOV T. Distributed representations of sentences and documents[C]// 31st International Conference on Machine Learning. 2014:1188-1196.

[7] TANG Z H,ZHU Q X,HONG C Q,et al. Based on self encoders and hypergraph learning[J]. Acta Automatica Sinica,2016,42(1):1014-1021. (in Chinese)
唐朝辉,朱清新,洪朝群,等. 基于自编码器及超图学习的多标签特征提取[J]. 自动化学报,2016,42(1):1014-1021.

[8] XING C,MA L,YANG X. Stacked denoise autoencoder based feature extraction and classification for hyperspectral images[J]. Journal of Sensors,2016(2016):1-10.

[9] HOU X,SHEN L,SUN K,et al. Deep feature consistent variational autoencoder[C]// 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). 2017:1133-1141.

[10] TAO C,PAN H B,LI Y S,et al. Unsupervised spectral-spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification[J]. IEEE Geoscience and Remote Sensing Letters,2015,12(12):2438-2442.

[11] CIREGAN D,MEIER U,SCHMIDHUBER J. Multi-column deep neural networks for image classification[C]// 2012 IEEE conference on Computer vision and pattern recognition (CVPR). 2012:3642-3649.

[12] KRIZHEVSKY A,SUTSKEVER I,HINTON G E. Imagenet classification with deep convolutional neural networks[C]// Advances in Neural Information Processing Systems. 2012:1097-1105.

[13] URIARTE-ARCIA A V,LÓPEZ-YÁNEZ I,YÁ NEZ-MÁRQUEZ C. One-hot vector hybrid associative classifier for medical data classification PloS one[J]. Public Library of Science,2014,9(10):95-105.

[14] ZHANG Y Y,HUO J,YANG W Q,et al. A deep belief network-based heterogeneous face verification method for the second-generation identity card[J]. CAAI Transactions on Intelligent Systems,2015,10(2):193-200. (in Chinese)
张媛媛,霍静,杨婉琪,等. others 深度信念网络的二代身份证异构人脸核实算法[J]. 智能系统学报,2015,10(2):193-200.

[15] HINTON G E,SALAKHUTDINOV R R. Replicated softmax:an undirected topic model[C]// 22nd International Conference on Neural Information Processing Systems. 2009:1607-1614.

[16] LV F,HAN M,QIU T. Remote Sensing Image Classification Based on Ensemble Extreme Learning Machine with Stacked Autoencoder[J]. IEEE Access,2017,3(99):1-11.

[17] GAO J,ZHANG C X,WANG Z,et al. Question Classification Based on Improved TFIDF Algorithm[C]// International Conference on Control,Automation and Artificial Intelligence. 2017:354-357.

[18] YANG B,HAN Q W,LEI M,et al. Short Text Classification Algorithm Based on Improved TF-IDF Weight[J]. Journal of Chongqing University of Technology(Natural Sicence),2016,30(12):103-113. (in Chinese)
杨彬,韩庆文,雷敏,等. 基于改进的 TF-IDF 权重的短文本分类算法[J]. 重庆理工大学学报(自然科学),2016,30(12):103-113.