文章编号:1007-5321(2020)05-0098-07 **DOI**:10.13190/j.jbupt.2020-033

一种基于高层特征融合的网络商品分类

刘逸琛, 孙华志, 马春梅, 姜丽芬, 钟长鸿
(天津师范大学 计算机与信息工程学院, 天津 300387)

摘要: 为了利用商品文本标题实现商品自动分类,提出一种基于高层特征融合的商品分类模型. 首先,提出基于字嵌入和词嵌入的文本底层特征表示法,进而获得更强的商品标题结构特征表达;其次,提出了联合自注意力、卷积神经网络和通道注意力的机制,对文本标题的底层特征进行增强并获得高层增强特征;最后,通过将文本的字嵌入和词嵌入的高层增强特征进行融合,最终获得商品文本标题的综合特征,并实现商品自动分类. 以商品标题语料作为数据集进行了实验,实验结果表明,该模型对三级商品类别的分类精度能够达到 84.348%,召回率和 F1 值分别达到了 47.8% 和 49.4%,优于现有可用于商品文本标题分类的先进短文本分类方法.

关 键 词: 商品分类; 短文本分类; 特征融合; 特征增强; 注意力机制
中图分类号: TP183 **文献标志码**: A

Commodity Classification of Online Based on High-Level Feature Fusion

LIU Yi-chen, SUN Hua-zhi, MA Chun-mei, JIANG Li-fen, ZHONG Chang-hong
(School of Computer and Information Engineering, Tianjin Normal University, Tianjin 300387, China)

Abstract: In order to realize automatic classification of commodities by leveraging text titles of commodities, a commodity classification model high-level feature fusion (HFF) based on high-level feature fusion is proposed. Firstly, a char embedding and word embedding based low-level feature representation method for the text title is proposed. Then a stronger feature expression of the commodity title structure can be obtained. Secondly, a joint self-attention mechanism, convolutional neural network, and channel attention are proposed to enhance the low-level features and obtain high-level enhancement features of the text title. Finally, by fusing the high-level enhancement features of the word embedding and the char embedding of the text, a comprehensive feature of the text title of the commodity is finally obtained and used for the commodity classification. Experiments are conduct on the dataset of the commodity titles. The experiments show that the classification accuracy of HFF for the third-level commodity can reach 84.348%. In addition, the recall and the F1 value of the HFF reach 47.8% and 49.4%, respectively, which is superior to the existing advanced short text classification method that can be used for the commodity text titles classification.

Key words: commodity classification; short text classification; feature fusion; feature enhancement; attention

收稿日期: 2020-04-23
基金项目: 国家自然科学基金项目(61702370); 天津市自然科学基金项目(18JCYBJC85900, 18JCQNJC70200); 天津市科技发展战略研究计划项目(17ZLZXZF00530); 天津市教委科研计划项目(JW1702)
作者简介: 刘逸琛(1995—), 男, 硕士生.
通信作者: 马春梅(1985—), 女, 讲师, E-mail:mcmxhd@163.com.

随着电商平台的迅速崛起，网络零售逐渐成为当下热门的商品交易方式之一，而快速、准确地对所出售的商品进行智能分类显得尤为重要. 一方面，赋予商品适当的类别标志，即将商品自动地逐级划分，能够方便消费者选择性购买；另一方面，能够准确识别隐藏在交易背后的违规产品，规范网络零售交易环境. 若有发布或售卖国家规定的违禁药品、管制刀具等违禁产品的情况，通过商品自动分类技术能将此类商品自动划分为违规类别，禁止商品发布或自动将已发布商品下架.

尽管电商平台大多具有商品分类的功能，但大多是通过网站编辑或网络卖家进行人工分类，对商品自动化分类的相关研究也较少. 商品的文本标题通常包含商品大部分特征信息，是描述商品属性较为全面的信息源. 根据对网络零售平台商品标题分类数据集的分析，笔者发现商品标题具有以下 3 个特点.

1）字符数量少. 由于标题字数最长不超过 100 个字，平均字数为 43 个字，少于传统意义上的短文本字数.

2）文本长度跨度大. 对数据集中商品文本标题长度的统计结果如图 1 所示. 商品标题的长度分布呈正态分布趋势，最短的标题为 5 个字，最长的标题为 97 个字，这种字数差距为分类模型的建立带来了更大的挑战.

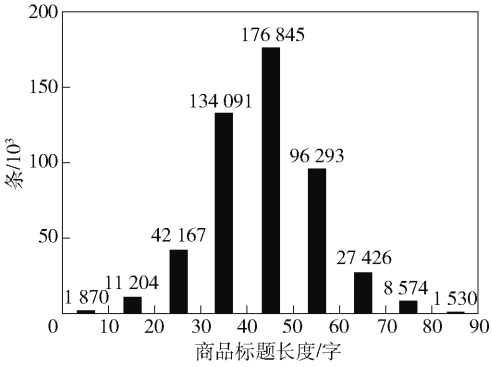

图 1　商品文本标题长度统计

3）不遵循语法结构. 考虑到标题检索，卖家通常会将包含商品特征的关键词组成标题，所以标题信息仅由词语构成，基本不包含任何传统语法结构. 而且，虽然商品的标题文本字数少，但是包含的语义信息却很丰富.

这些特点都与传统短文本分类研究有很大不同. 另外，研究对象是商品文本标题信息，它与短文本的概念相似，都是较短的文本信息，但是二者的研究重点不同. 短文本的作用在于弥补因稀疏性而缺少的信息[1]，研究的重点在于对商品信息的特征进行抽取和识别，用其进行分类.

1　相关工作

商品文本标题的自动分类任务可以归结为短文本分类任务[2]，即为将给定的短文档分为 n 个类别中的一个或多个. 目前主要的方法有两大类：一类是传统文本分类方法；另一类是基于深度学习的文本分类方法.

传统的文本分类主要利用人工进行特征设计，然后利用分类器对其进行分类，如 Danesh[3] 等使用决策树算法对特定短文本进行分类. 但传统方法的文本表示高维度高稀疏，特征表达能力较弱，严重影响了分类精度.

深度神经网络能够通过深层次网络结构自动挖掘数据细粒度特征，是最成功的机器学习方法之一. 对于短文本分类问题，在深度学习方面，一些学者提出利用卷积神经网络（CNN，convolutional neural networks）或循环神经网络（RNN，recurrent neural network）等网络结构自动获取短文本特征表达，代替繁杂的人工特征工程，通过端到端的方式对文本进行分类. 例如，Joao 等[4] 提出的 TextCNN（convolutional neural network for the classification of sentences）利用 CNN 模型提取句子中文本局部的相关性. CNN 模型的本质是提取区域特征，而自然语言处理中更多的是文本序列的问题，运用 RNN 能够更好地处理文本序列的时序关系. 因此，Xie 等[5] 使用了双向循环神经网络（Bi-directional RNN，bi-directional recurrent neural network）捕获变长双向的“n-gram”信息，并添加了自注意力机制，提升重要部分的权重，使文本分类精度有所提高. 由于循环神经网络偏重序列中的近期输入，无法提取语句中距离相对较远的词语间的联系，而注意力机制通过对文本表示进行软对齐，构建文本表示之间的关系矩阵[6]，可以很好地解决此问题. Google Brain 使用自注意力机制来学习文本表示[7].

尽管上述方法在短文本分类中已经取得了一定的效果，但由于商品标题的特点，当前方法仍然存在以下问题：① 词向量获取信息不够全面，信息源单一，不能体现标题结构性的特点；② RNN 在商品文本标题信息中能够提取的时序信息较少，

且时间复杂度较高;③ 由于商品文本标题信息不遵循语法结构,现有模型无法有效地捕捉文本中词语间的关联.

2 高层特征融合模型

为了获取文本的细粒度特征,提高网络零售平台对商品信息分类的精度,提出了高层特征融合分类模型(HFF,high-level feature fusion),模型结构如图2所示. 网络结构分为嵌入输入模块、特征提取模块、特征增强模块和融合输出模块4个模块. 通过嵌入输入模块的词嵌入(word embedding)和字嵌入(char embedding)能够将商品文本标题的字向量和词向量映射成为其底层输入特征;特征提取模块使用双层多卷积核的卷积神经网络对输入的底层特征进行特征提取,形成辨识度更高的高层特征;为了增强2种特征表达能力,抓取特征之间的全局依赖关系,特征增强模块使用自注意力(self-attention)和通道注意力(channel attention)2种机制分别对文本的底层特征和高层特征进行增强;融合输出模块将提取到的字、词两方面的特征拼接后通过线性变换改变特征向量维度,最终通过输出文本的概率分布,获得商品类别.

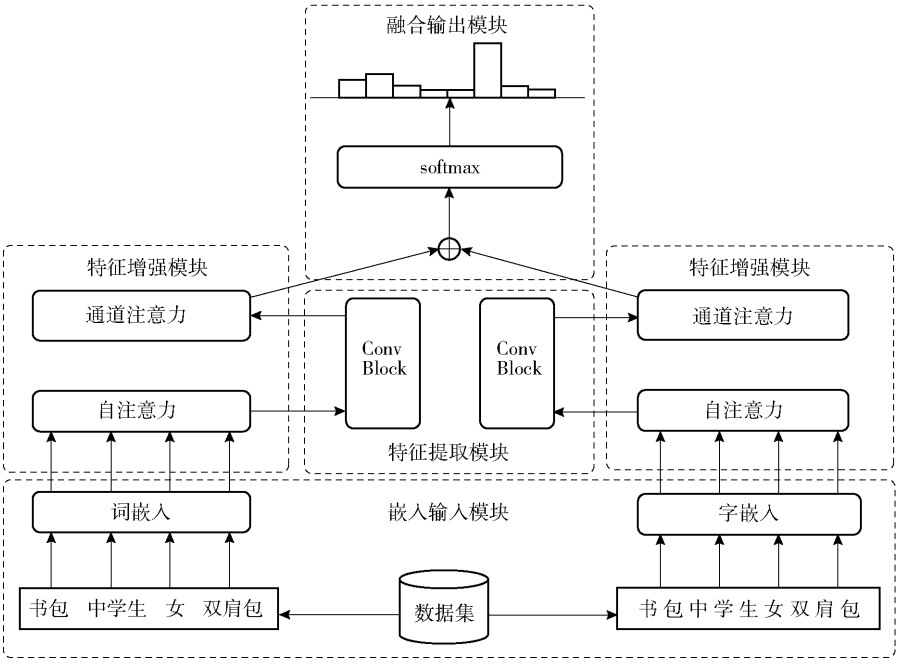

图2 高层特征融合模型结构

2.1 多特征的嵌入输入

在商品文本标题分类的任务中,存在字符数量少、文本长度跨度大和不遵循语法结构等特点,且词之间具有较强的结构性,单独的嵌入方式不能获取足够的文本信息. 因此,在嵌入输入模块中基于字嵌入和词嵌入的2种嵌入方式,能够分别将文本标题的字序列和词序列映射到可计算的高维特征向量空间,形成文本底层特征.

对于原始文本 T,为了获得HFF模型的字序列和词序列输入,通过北京大学的开源分词工具pkuseg[8]对输入文本进行了分词,进而获得字序列 $C=\{c_1,c_2,\cdots,c_q\}$ 和词序列 $W=\{w_1,w_2,\cdots,w_p\}$,其中 q 和 p 分别为文本 T 的字与词语序列的固定长度. 对于较长的序列长度,将超过固定长度的部分删除;对于不足序列长度的部分,将使用在字典中设定的特定空缺符号进行填充.

为了获得语义关联更丰富的字序列和词序列的表达,提高模型底层特征数据质量,以维基百科中文语料为词向量生成语料库,通过分词工具pkuseg对其进行分词,然后利用文本嵌入模型[9](word2vec)获得该语料的字向量和词向量字典. 通过查找字典进而获得商品标题文本的字序列元素 c_i 和词序列元素 w_j 的向量表示,即字向量 $\boldsymbol{x}_{c_i}$ 和词向量 $\boldsymbol{x}_{w_j}$. 由此,文本 T 的底层输入特征可以表示为 $\boldsymbol{X}^C=(\boldsymbol{x}_{c_1},\boldsymbol{x}_{c_2},\cdots,\boldsymbol{x}_{c_q})$ 和 $\boldsymbol{X}^W=(\boldsymbol{x}_{w_1},\boldsymbol{x}_{w_2},\cdots,\boldsymbol{x}_{w_p})$.

2.2 高层特征提取

高层特征表达旨在通过将不同尺度的文本信息组合在一起,生成具有丰富语义特征、对文本区分度

更高的高层特征表示. 针对商品文本标题的特点，在特征提取模块选用双层卷积网络用于特征提取，其结构如图 3 所示. 为了捕获不同尺度下文本的高层特征，该模块在每层卷积过程中分别使用 3 种尺度的卷积核(e_1，e_2，e_3)，其中 embed 表示嵌入维度，Conv1d 表示一维卷积操作. 利用各个卷积核卷积后，通过激活函数 ReLU 的作用，最终经过两层卷积神经网络，将得到的 3 个特征矩阵经过最大池化之后，在最低维度进行拼接，最终获得文本的高层特征. 具体而言，对于经过自注意增强后的底层特征 $\boldsymbol{X}'^n$，在卷积核 $e_l \in \{e_1, e_2, e_3\}$ 的作用下，经过双层卷积神经网络后，获得的特征向量表示为

$$\boldsymbol{Y}_{e_l}^n = P_{\max}[f(\boldsymbol{\omega}\boldsymbol{X}'^n + \boldsymbol{b})] \tag{1}$$

其中：e_l 为第 l 个卷积核；$P_{\max}$ 为最大池化操作；f 为 ReLU 激活函数；$\boldsymbol{\omega}$ 为卷积核参数矩阵；$n \in \{C, W\}$；$\boldsymbol{b}$ 为偏置量. 因此，经过特征拼接后获得的字序列或者词序列的高层特征为

$$\boldsymbol{Y}^n = \boldsymbol{Y}_{e_1}^n \oplus \boldsymbol{Y}_{e_2}^n \oplus \boldsymbol{Y}_{e_3}^n \tag{2}$$

其中⊕表示向量在最低维度的拼接.

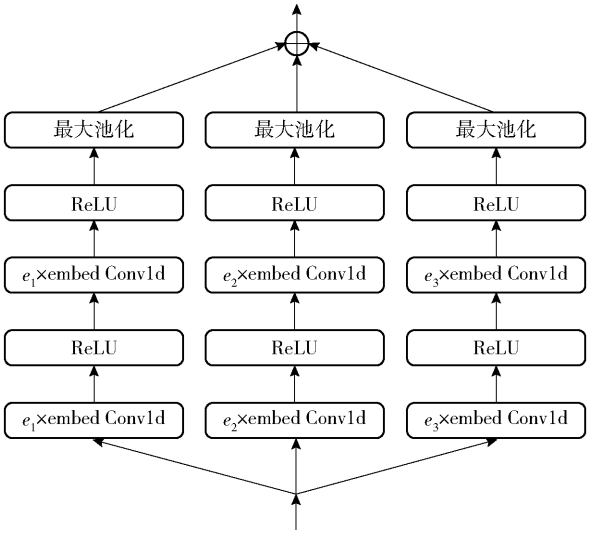

图 3　特征提取模块结构

2.3　特征增强

为了增强特征表达能力，提高文本分类的结果，在特征增强模块中提出了基于 2 种不同注意力机制的特征增强方法，分别用于底层特征和高层特征的特征增强.

对于底层特征，利用自注意力机制增强特征表达的方法，使其更加注重局部特征的突出，在一定程度上弥补文本信息不足的缺陷；对于高层特征，在模型中引入通道注意力机制对高层特征进行增强，使模型更加注重文本全局特征的增强，进而解决商品标题文本长度跨度大的问题. 两者的共同作用可以使模型从有限的特征中提取到更有效的信息，进而获得更好的分类效果.

2.3.1　底层特征增强

自注意力机制通过计算文本序列中单词“重要性”的分布，进而获得局部增强特征. 具体而言，首先，底层特征向量 $\boldsymbol{X}^n$ 要经过线性变换，变换后的特征可以表示为

$$\widetilde{\boldsymbol{X}}^n = \widetilde{\boldsymbol{\omega}}\boldsymbol{X}^n + \widetilde{\boldsymbol{b}} \tag{3}$$

其中 $\widetilde{\boldsymbol{\omega}}$ 和 $\widetilde{\boldsymbol{b}}$ 均为线性变换中的参数矩阵；其次，用变换后的特征同其转置相乘，并通过 softmax 函数可以获得特征的权重矩阵 $\boldsymbol{A}_n$，表示为

$$\boldsymbol{A}_n = \mathrm{softmax}[\widetilde{\boldsymbol{X}}^n \boldsymbol{W}^{Xn} (\widetilde{\boldsymbol{X}}^n)^{\mathrm{T}}] \tag{4}$$

其中：$\boldsymbol{W}^{Xn}$ 为自注意力机制中要学习的权重矩阵；softmax(·)表示将输入通过一个归一化指数函数进行处理，可将输出向量压缩至[0，1]之间，且保证所有元素和为 1. 令 $\boldsymbol{Z} = \widetilde{\boldsymbol{X}}^n \boldsymbol{W}^{Xn} (\widetilde{\boldsymbol{X}}^n)^{\mathrm{T}}$，则 $\boldsymbol{A}_n$ 中第 j 个元素的计算公式为

$$\boldsymbol{a}_j = \frac{\exp(\boldsymbol{Z}_j)}{\sum_k \exp(\boldsymbol{Z}_k)} \tag{5}$$

其中 $\boldsymbol{Z}_j$ 和 $\boldsymbol{Z}_k$ 为 $\boldsymbol{Z}$ 中第 j 和 k 个元素. 因此，增强后的特征为

$$\boldsymbol{X}'^n = \boldsymbol{A}_n \widetilde{\boldsymbol{X}}^n \tag{6}$$

2.3.2　高层特征增强

高层特征增强模块位于特征提取模块之后，它通过整合所有通道中的相关特征，有选择地增强相互关联的通道特征图，以此来抓取特征之间的全局依赖关系，并对提取的特征进行增强，其过程主要分为压缩和扩展 2 个阶段. 压缩阶段用于全局特征的集中表述，旨在使用少量特征表述出文本中最重要的信息. 该过程先通过执行单维全局平均池化，经过特征提取模块得到高层特征向量 $\boldsymbol{Y}^C$ 和 $\boldsymbol{Y}^W$，通过全局平均池化操作，将最后一维特征向量压缩为 1，使其可以代表一定范围内的特征；再使用一个一维卷积操作，并通过 ReLU 函数激活，将矩阵维度 $BE \times 1$ 压缩到 $B(E/R) \times 1$，其中 B 为一次训练所选取的样本数，E 为特征通道数，R 为一个固定值，用于按照固定比例压缩通道数. 经过通道特征压缩后获得的增强特征 $\boldsymbol{S}^n$ 可以表示为

$$\boldsymbol{S}^n = f\left[\boldsymbol{W}_1^{S^n} H_{\mathrm{GP}}(\boldsymbol{Y}^n)\right] = f\left(\boldsymbol{W}_1^{S^n} \frac{1}{r}\sum_{i=1}^{r} \boldsymbol{Y}_i^n\right) \quad (7)$$

其中:$f(\cdot)$为 RuLU 函数,$\boldsymbol{W}_1^{S^n}$ 为压缩阶段卷积操作的参数矩阵,$r \in \{q, p\}$ 为字嵌入或词嵌入维度,$H_{\mathrm{GP}}(\cdot)$为全局平均池化操作,$\boldsymbol{Y}^n \in \{\boldsymbol{Y}^C, \boldsymbol{Y}^W\}$ 为待压缩的通道特征.

扩展阶段用于为每个特征图生成通道加权,通过对压缩后获得的增强特征进行扩展,学习权重矩阵以建模通道特征图之间的相关性,更好地拟合通道间的关联. 为此,使用一维卷积操作并通过 Sigmoid 函数运算,将矩阵维度由压缩之后的 $B(E/R)\times 1$ 变回 $BE\times 1$,经扩展,通道权重向量 $\boldsymbol{A}_n'$ 可以表示为

$$\boldsymbol{A}_n' = \delta(\boldsymbol{W}_2^{S^n}\boldsymbol{S}^n) \quad (8)$$

其中:$\delta(\cdot)$为 Sigmoid 函数,用于将提取特征映射至[0,1]之间,作为权重值;$\boldsymbol{W}_2^{S^n}$ 为扩展阶段卷积操作的参数矩阵.

通道注意力机制的最终输出是经过加权的特征矩阵 $\boldsymbol{Y'}^C$ 和 $\boldsymbol{Y'}^W$,可以表示为

$$\boldsymbol{Y'}^n = \boldsymbol{A}_n'\boldsymbol{Y}^n \quad (9)$$

2.4 融合输出

令特征增强后字嵌入和词嵌入的高层特征分别为 $\boldsymbol{Y'}^C$ 和 $\boldsymbol{Y'}^W$. 为了获得对文本表达更强的特征,以提高分类精度,构建了融合输出模块,通过将 2 个特征向量在低维度上进行拼接,经过一个线性变换过程后被 softmax 函数激活. 因此,具有标题文本 T 的商品类别预测概率为

$$P = \mathrm{softmax}[\boldsymbol{W}^L(\boldsymbol{Y'}^C \oplus \boldsymbol{Y'}^W) + \boldsymbol{b}^L] \quad (10)$$

其中:$\oplus$为特征拼接操作,$\boldsymbol{W}^L$ 为全连接层的参数矩阵,$\boldsymbol{b}^L$ 为全连接层的偏执量.

为了训练模型,构建二元交叉熵[10]损失函数作为模型目标函数,如式(11)所示. 它表示商品预测类别与其真实类别的接近程度,交叉熵值越大,预测类别同真实类别的近似程度越低;反之亦然. 为了学习模型参数,采用反向传播机制对高层特征融合模型进行训练和更新,以最小化商品真实类别和商品预测类别的交叉熵损失,损失函数为

$$l = -\sum\sum c_i^{j} \lg(c_i^{j}) \quad (11)$$

其中:c 为该商品的真实类别,c'为所提出模型的预测类别,i 为商品索引,j 为商品类别索引.

3 实验与性能分析

3.1 实验设置

3.1.1 数据集

使用网络零售平台商品标题分类数据集[11]进行实验,此数据集共包含了 50 万条商品数据,通常商品标题由用户设定,因此该数据集对商品文本标题的结构特点具有泛化性. 该数据集分为 30 个一级类别,192 个二级类别和 1 258 个三级类别. 按照类内随机划分的机制,将数据集划分为训练集和测试集,其中训练集样本有 35 万条,占样本总数的 70%,测试集样本有 15 万条,占样本总数的 30%.

3.1.2 实验环境和参数设置

实验选用已经广泛用于深度学习模型实现和训练的 pytorch[12]框架,在配备 NVIDIA GTX 1080Ti GPU 和 Intel Xeon E5-2620 CPU 的计算机上进行训练和测试.

网络结构中的部分参数根据实验环境设置,如每批次数量(batch-size)和优化器的选择. 部分超参数是通过对数据集的分析而决定的,模型具体参数如表 1 所示. 在评价指标方面,采用了准确率、召回率和 F1 测量值对模型的分类能力进行分析.

表 1 实验参数设置

参数	值
字嵌入维度/维	50
词嵌入维度/维	200
卷积核大小/维	3,4,5
批处理大小/个	64
字序列最大长度/字	60
词序列最大长度/词	30

3.2 实验结果

3.2.1 特征长度分析

特征长度是在进行底层特征表示时所使用字或词序列的长度. 如果输入文本的长度大于特征长度,则舍弃超出长度的特征信息. 如果输入文本的长度小于特征长度,则使用在字典中设定的特定空缺符号进行填充. 由于商品文本标题长度跨度比较大,选取不同特征长度会对实验效果产生不同的影响. 为此,根据图 1 描述的商品标题长度分布,在实验中分别设置了词序列和字序列长度值为“20,40”、“30,60”、“40,80”3 组参数. 在 3 组不同的特征长度下,HFF 模型的实验结果如表 2 所示. 可以

看出，当字和词的特征长度为“30，60”时，HFF 性能最优. 因此，分别设置在 HFF 模型中字嵌入和词嵌入的序列长度为 30 和 60.

表 2　不同特征长度下三级商品类别的分类实验结果

词序列长度/词	字序列长度/字	准确率/%	召回率/%	F1 测量值
20	40	84.231	47.50	0.490
30	60	84.348	47.80	0.494
40	80	83.331	47.70	0.492

3.2.2　实验对比分析

在实验中，分别对二级类别和三级类别的商品进行了分类预测. 此外，为了验证 HFF 模型的有效性，还与 CharCNN[4]、WordCNN[4]、SABiLSTM（Self-Attention-based BiLSTM）[5]、CRAN[13]、Transformer[7]、RInception（BLSTM-Inception）[14] 和 DeepCNN[4]7 种短文本分类模型的性能进行了实验对比. 其中 CharCNN、WordCNN 和 DeepCNN 均为对文献[4]中模型复现的针对中文文本的改进版本，其区别在于 CharCNN 使用字嵌入，WordCNN 使用词嵌入，DeepCNN 为在 WordCNN 基础上将模型扩展到 6 层 CNN 结构，并且在中间层使用了残差连接增强训练效果；Transformer 为文献[7]中模型的编码器后接一个线性层分类模型；RInception 为文献[14]中 BLSTM-Inception 模型的缩写，其中 R 表示模型中 BLSTM 的编码部分.

不同方法对于二级商品类别和三级商品类别的分类结果如表 3 和表 4 所示. 其中，HFF-Char、HFF-Word 和 HFF-Attention 分别为 HFF 模型的字嵌入、词嵌入和注意力机制消解实验模型.

表 3　二级商品类别的分类实验结果

模型	准确率/%	召回率/%	F1 测量值
CharCNN	91.293	72.20	0.716
WordCNN	93.334	74.80	0.750
SABiLSTM	93.379	75.20	0.753
CRAN	93.551	75.60	0.763
Transformer	91.263	74.20	0.745
RInception	91.648	75.10	0.754
DeepCNN	92.589	74.70	0.749
HFF	93.756	76.70	0.769
HFF-Char	93.134	75.90	0.761
HFF-Word	90.103	74.30	0.745
HFF-Attention	93.431	76.00	0.763

表 4　三级商品类别的分类实验结果

模型	准确率/%	召回率/%	F1 测量值
CharCNN	82.765	21.20	0.200
WordCNN	83.496	46.00	0.490
SABiLSTM	83.341	46.60	0.480
CRAN	83.934	44.60	0.433
Transformer	75.82	45.60	0.473
RInception	79.831	44.60	0.470
DeepCNN	78.11	43.40	0.468
HFF	84.348	47.80	0.494
HFF-Char	83.786	47.00	0.486
HFF-Word	82.945	46.30	0.480
HFF-Attention	83.958	46.60	0.473

由表 3 和表 4 中可见，HFF 模型对二级商品类别和三级商品类别的分类精度、召回率和 F1 值分别达到了 93.756%、76.70%、0.769 和 84.348%、47.80%、0.494，均高于实验中其他 7 种基线模型. 在这些方法中，由于商品文本标题的短缺性和稀疏性导致其无法包含足够丰富的信息，致使卷积神经网络不能有效捕获其中更深层特征，RInception 和 DeepCNN 这 2 种深层卷积神经网络模型在上述 2 种级别分类结果中的效果最差. 因此，尽管深层卷积神经网络模型适用于文档级数据集，但它在解决依据商品短文本标题进行分类问题方面的效果有限. 另外，在实验中发现，仅使用词嵌入的卷积神经网络模型其准确率已经超过 SABiLSTM 模型的准确率. 由此可见，在不遵循语法结构的短文本分类中，卷积神经网络相对于循环神经网络更有优势.

分别以词组、字符为特征的 CNN 实验结果表明，以词组为特征的 CNN（word CNN）性能好于以字符为特征的 CNN（Char CNN），说明词组在编码过程中包含了比字符更多的信息. 但是由于仅使用字符作为 CNN 的输入也可以使准确率达到 82.765%，说明字符嵌入也同样包含大量的信息.

首先，由于 2 种特征嵌入的组合包含了更多的信息，在一定程度上能够缓解文本短缺及稀疏性的问题；其次，引入了自注意力和通道注意力，增强了特征在空间和全局依赖关系式上的表达，解决了商品标题文本长度跨度大的问题；最后，使用卷积神经网络提取特征，对于不遵循语法结构的短文本数据具有更好的效果，因此 HFF 模型的精度高于其他方

法. 通过对字嵌入和词嵌入的消解实验结果对比,使用单一嵌入方式的准确率均不及2种嵌入方式融合的准确率,由此可以证明模型能够从2种不同的嵌入方式中获取更多有效的信息. 另外,注意力机制消解实验结果对比证明,在模型中引入注意力机制可以有效地提升分类精度.

模型收敛速度也是验证模型实用性的一个非常重要的指标,为此,在实验中使用一块GTX1080ti显卡,在网络零售平台商品标题分类数据集上进行三级商品类别的分类,以测试各个模型的收敛情况. 图4所示为HFF模型与其他7种方法的收敛效果. 可以看出,提出的HFF模型的收敛较为平稳且精度更高,各个模型的收敛时间如表5所示. 由图4和表5可见,HFF模型的收敛时间相较于普通的CNN模型和添加了RNN结构的模型其收敛速度更快. 但是由于特征提取模块由双层多卷积核卷积神经网络构成,所以收敛时间弱于Transformer模型和RInception模型. 尽管Transformer和RInception模型的收敛速度快,但是其分类效果较差. 因此,提出的HFF模型不仅分类准确率更高,收敛速度相对来说也较快,同时收敛过程更加稳定,充分证明了模型的有效性和优越性.

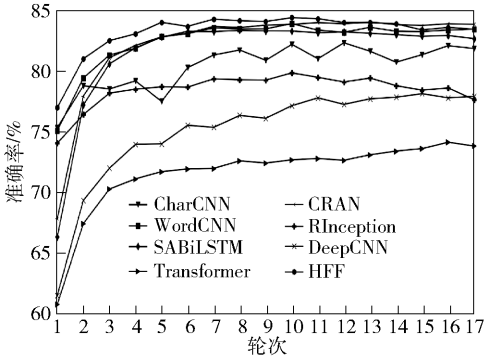

图4 三级商品类别的分类准确率对比

表5 三级商品类别模型的收敛时间

模型	收敛时间/s	模型	收敛时间/s
CharCNN	907	Transformer	363
WordCNN	730	RInception	326
SABiLSTM	494	DeepCNN	613
CRAN	650	HFF	480

4 结束语

通过分析商品文本标题的特点可知,商品文本标题具有字符数量少、文本长度跨度大和不遵循语法结构等特点,因此提出将一种高层特征融合模型用于商品分类的方案. 在该模型中,从字嵌入和词嵌入两方面对商品文本的信息进行表示,并用自注意力机制对2种特征进行增强,在一定程度上解决了文本信息不足的问题. 此外,为了解决商品标题文本长度跨度大的问题,提出了基于两阶段,即压缩和扩展的高层特征增强方法,增加了通道间关联性. 最后,通过实验验证得出结论:对于不遵循语法结构的短文本数据,卷积神经网络具有更好的效果. 通过在网络零售平台商品标题分类数据集上进行的对比实验结果验证了HFF模型在综合商品分类的性能和模型收敛时间上优于其他7种短文本分类模型.

参考文献:

[1] Xu Jingyun, Cai Yi. Incorporating context-relevant knowledge into convolutional neural networks for short text classification[C] //Proc of the 33rd AAAI Conf on Artificial Intelligence. Honolulu: AAAI, 2019: 10067-10068.

[2] Ma Chenglong, Xu Weiqun, Li Peijia, et al. Distributional representations of words for short text classification [C] //Proc of the 51nd Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL). Stroudsburg: ACL, 2015: 33-38.

[3] Danesh I, Steve W, Calton P. Study of static classification of social spam profiles in myspace[C] //Proceedings of the 4th Int AAAI Conf on Weblogs and Social Media (ICWSM). Menlo Park: AAAI, 2010: 70-75.

[4] Joao P A V, Raimundo S M. An analysis of convolutional neural networks for sentence classification[C] //Proceedings of Conferencia Latinoamericana En Informatica (CLEI). Piscataway: IEEE, 2017: 1-5.

[5] Xie Jun, Chen Bo, Gu Xinglong, et al. Self-attention-based bilstm model for short text fine-grained sentiment classification [J]. IEEE Access, 2019, 7: 180558-180570.

[6] 朱皓, 谭咏梅. 基于胶囊的英文文本蕴含识别方法[J]. 北京邮电大学学报, 2019, 42(3): 21-28.
Zhu Hao, Tan Yongmei. English textual entailment recognition using capsules[J]. Journal of Beijing University of Posts and Telecommunications, 2019, 42(3): 21-28.

(下转第117页)

[11] Gupta H, Gupta D. LPC and LPCC method of feature extraction in speech recognition system[C]//2016 6th International Conference Cloud System and Big Data Engineering. Noida: IEEE Press, 2016: 498-502.

[12] Eltiraifi O, Elbasheer E, Nawari M. A comparative study of MFCC and LPCC features for speech activity detection using deep belief network[C]//2018 International Conference on Computer Control, Electrical and Electronics Engineering. Khartoum: IEEE Press, 2018.

[13] Liu Gang, He Wei, Jin Bicheng. Feature fusion of speech emotion recognition based on deep learning[C]//2018 International Conference on Network Infrastructure and Digital Content. Guiyang: IEEE Press, 2018: 193-197.

[14] Nikhitha M, Roopa Sir S, Uma Maheswari B, et al. Fruit recognition and grade of disease detection using inception V3 model[C]//2019 3rd International Conference on Electronics, Communication and Aerospace Technology. Coimbatore: IEEE Press, 2019: 1040-1043.

(上接第 104 页)

[7] Ashish V, Noam S, Niki P, et al. Attention is all you need[C]//Proc of the 30th Advances in Neural Information Processing Systems(NIPS). Cambridge: MIT Press, 2017: 6000-6010.

[8] Luo Ruixuan, Xu Jingjing, Zhang Yi, et al. PKUSEG: a toolkit for multi-domain Chinese word segmentation[EB/OL]. 2019(2019-06-27)[2020-07-18]. https://arxiv.org/abs/1906.11455.

[9] Tomas M, Kai C, Greg C, et al. Efficient estimation of word representations in vector space[C]//Proc of the 1rd Int Conf on Learning Representations(ICLR). Scottsdale: ICLR, 2013: 1-12.

[10] Reuven Y. Rubinstein. Optimization of computer simulation models with rare events[J]. European Journal of Operational Research, 1997, 99(1): 89-112.

[11] GitHub. DFF-Dataset[EB/OL]. 2020(2020-06-18)[2020-10-04]. https://github.com/liuyichenaa1/DFF-Dataset.

[12] Adam P, Sam G, Francisco M, et al. pyTorch: an imperative style, high-performance deep learning library[C]//Proc of the 32th Advances in Neural Information Processing Systems(NeurIPS 2019). Vancouver: MIT Press, 2019: 8024-8035.

[13] Guo Long, Zhang Dongxiang, Wang Lei, et al. CRAN: a hybrid CNN-RNN attention-based model for text classification[C]//Proc of the 37th International Conference on Conceptual Modeling(ER2018). Berlin: Springer-Verlag, 2018: 571-585.

[14] Jiang Wei, Jin Zhong. Integrating bidirectional LSTM with inception for text classification[C]//Proc of the 4th IAPR Asian Conference on Pattern Recognition(ACPR 2017). Piscataway: IEEE, 2017: 870-875.