doi:10.6041/j.issn.1000-1298.2021.07.025

基于 Attention_DenseCNN 的水稻问答系统问句分类

王郝日钦[1,2]　吴华瑞[3,4]　冯　帅[1]　刘志超[1]　许童羽[1]

(1.沈阳农业大学信息与电气工程学院,沈阳 110866;2.内蒙古民族大学计算机科学与技术学院,通辽 028043;
3.国家农业信息化工程技术研究中心,北京 100097;4.北京农业信息技术研究中心,北京 100097)

摘要:为了解决“中国农技推广 APP”问答社区中水稻提问数据快速自动分类的问题,提出一种基于 Attention_DenseCNN 的水稻文本分类方法。根据水稻文本具备的特征,采用 Word2vec 方法对文本数据进行处理与分析,并结合农业分词词典对文本数据进行向量化处理,采用 Word2vec 方法能够有效地解决文本的高维性和稀疏性问题。对卷积神经网络(CNN)上下游卷积块之间建立一条稠密的链接,并结合注意力机制(Attention),使文本中的关键词特征得以充分体现,使文本分类模型具有更好的文本特征提取精度,从而提高了分类精确率。试验表明:基于 Attention_DenseCNN 的水稻问句分类模型可以提高文本特征的利用率、减少特征丢失,能够快速、准确地对水稻问句文本进行自动分类,其分类精确率及 F1 值分别为 95.6% 和 94.9%,与其他 7 种神经网络问句分类方法相比,分类效果明显提升。

关键词:水稻问句分类;自然语言处理;密集连接卷积神经网络;注意力机制

中图分类号:TP183　　**文献标识码**:A　　**文章编号**:1000-1298(2021)07-0237-07　　**OSID**:

Classification Technology of Rice Questions in Question Answer System Based on Attention_DenseCNN

WANG Haoriqin[1,2]　WU Huarui[3,4]　FENG Shuai[1]　LIU Zhichao[1]　XU Tongyu[1]

(1. *College of Information and Electrical Engineering*, *Shenyang Agricultural University*, *Shenyang* 110866, *China*
2. *College of Computer Science and Technology*, *Inner Mongolia University for Nationalities*, *Tongliao* 028043, *China*
3. *National Engineering Research Center for Information Technology in Agriculture*, *Beijing* 100097, *China*
4. *Beijing Research Center for Information Technology in Agriculture*, *Beijing* 100097, *China*)

Abstract: In the QA community of Chinese Agricultural Technology Promotion APP, thousands of rice text data questions are added every day, and the rapid and automatic classification of questions is a key step to realize the intelligent QA system of rice. However, due to the high dimensional sparsity of text data and the particularity of agricultural problems, the classification of rice questions faces difficult challenges. In order to improve the classification performance of rice question text, a convolution text classification method with dense connection was proposed. A dense connection between upstream and downstream convolution blocks was established, which enabled the model to synthesize large-scale features from small-scale features. Combined with the agricultural word segmentation dictionary, the text data was segmented into 100-dimensional word vectors by Word2vec. Neural network model's parameters for question classification in rice question answering system were obtained by training text data with dense concatenated convolution model and attention mechanism. The experimental results showed that the text classification model based on Attention_DenseCNN can optimize the text's representation and feature extraction, and also it can automatically classify the rice question text with accuracy of 95.6% and F1 value of 94.9%. Compared with the other seven text classification methods, the classification performance had obvious advantages.

Key words: rice question classification; natural language processing; dense connection convolution neural network; attention mechanism

收稿日期:2020-09-19　修回日期:2020-10-28

基金项目:国家重点研发计划项目(2018YFD0300309)、江苏大学农业装备学部项目和内蒙古民族大学科学研究基金项目(NMDYB18028、NMDYB18026、NMDYB17138)

作者简介:王郝日钦(1988—),男,博士生,内蒙古民族大学实验师,主要从事农业智能系统与自然语言处理研究,E-mail:haoriqin@imun.edu.cn

通信作者:许童羽(1967—),男,教授,博士生导师,主要从事农业智能系统研究,E-mail:yatongmu@163.com

0 引言

“中国农技推广 APP”是一个提供专家指导、农技问答的综合性专业服务平台,用户每天在水稻问答模块中提问千余条,目前主要利用人工筛选特征和浅层学习模型对水稻提问数据进行分类管理。由于水稻提问数据的高维性和稀疏性,其分类效果并不理想,如何使用自然语言处理[1]技术自动、快速地挖掘水稻问句的特征,并用于水稻问句分类,是当前需要解决的重要难题。问答系统一般包括用户问句分类[2]、问句语义分析[3]和答案抽取[4]3 部分。问句分类是问答系统的关键模块,其分类准确性对问答系统至关重要,因此,对水稻问句进行自动分类是实现水稻智能问答系统[5]的关键,也是实现人工智能[6]的必然需求。

近年来,国内外研究人员使用机器学习[7-8]与深度学习[9-13]模型进行文本分类,并取得了良好的效果。卷积神经网络在图像分类领域取得了显著的效果,研究人员将卷积神经网络模型进行优化和改进,使用词向量转换工具将高维的文本数据转换为低维稠密的词向量,并输入卷积神经网络,提升了分类效果。文献[14]提出用于文本分类的单层卷积神经网络模型,将文本当作固定长度的图像输入到模型中,提高了文本分类效果,但并未考虑文本信息中的序列信息所带来的影响。文献[15]运用循环神经网络模型(Recurrent neural network,RNN)针对文本中的序列信息提升了文本分类效果,但 RNN 无法解决长时依赖以及梯度消失问题。为此,研究人员对 RNN 模型进行了优化及改进,提出了长短期记忆网络(LSTM)和门控循环单元神经网络模型(GRU),并用来解决文本分类问题。文献[16]提出 Densenet 网络模型,在卷积层之间采用了密集的连接,缓解了梯度消失问题,加强了特征传播,在文本分类任务上表现出良好的性能。在机器学习中,常用的文本分类方法有 K-means 算法[17]、朴素贝叶斯模型[18]和支持向量机模型[19]。在农业领域,由于缺少大规模可用的数据集,相关研究仍处于起步阶段。现有研究[20-23]为深度学习在农业文本数据分类提供了参考及可行性依据。然而,农业文本分类大部分采用固定宽度的卷积核来提取文本特征,未考虑关键词在农业文本分类中的重要性,在多类别水稻问答问句数据集上分类的验证也鲜有涉及。

针对水稻问答问句稀疏性强、噪声大、数据量大的特点,本文提出一种基于注意力机制的密集连接卷积模型文本分类方法。在 CNN 上下游卷积块之间建立了一条稠密的捷径,每个卷积块采用不同宽度的卷积核来提取文本特征,引入农业词典对文本数据进行分词和词向量转换,然后输入到建立的混合神经网络模型中,从而使网络可以自动提取水稻文本特征,并进行分类,针对一些重要参数进行优化和改进,以实现水稻问句的自动分类。

1 文本预处理及模型构建

1.1 语料库构建

本文数据为从“中国农技推广”APP 问答社区中导出的水稻提问数据,使用 Python 中的正则表达式对所获得的文本数据进行清洗和过滤,去除掉无用信息,对水稻提问数据进行人工标注,获得水稻病虫害、草害药害、高产增产、储存保鲜、栽培管理、其他共 6 个类别 12 000 条提问数据。水稻提问文本数据样例如表 1 所示。

表 1 水稻提问文本数据

Tab.1 Example of rice question

编号	问句	类别
1	水稻插秧什么样的密度既高产,米质又好?	高产增产
2	水稻细菌性条斑病怎么防治?	病虫害
3	南方水稻秧苗防寒保温技术要点有哪些?	储存保鲜
4	目前主推超级稻品种有哪些?	其他
5	什么是水稻二系杂交育种?	栽培管理
6	水稻田三棱草如何防治?	草害药害
7	水稻稻瘟病怎样识别和防治?	病虫害

1.2 数据预处理

首先对文本数据的常用标点符号进行去除,清洗和过滤无用信息。然后加载农业专用停用词表,去除文本中的常用停用词、中英文特殊字符等影响特征提取的噪声。采用 Python 的 JIEBA 分词工具包对文本数据进行分词,由于农业文本数据具有较多的农业专用词汇,为了提高分词的准确度,选择搜狗农业词汇大全[24]作为分词字典。预处理前后文本数据如表 2 所示。

表 2 预处理前后文本数据

Tab.2 Text data before and after pretreatment

数据预处理前	数据预处理后
水稻恶苗病的危害症状有哪些?	水稻/恶苗病/危害症状
水稻秧苗出现黄叶是什么原因引起的?	水稻/秧苗/出现/黄叶/原因/引起
高产水稻插秧前后至足苗期间应怎样施肥?	高产/水稻/插秧/前后/足苗期间/施肥
水稻种子发芽的最适温度是多少度?	水稻/种子/发芽/最适/温度
水稻插植后的栽培管理技术主要是什么?	水稻/插植/后/栽培管理/技术

1.3 数据向量化

由于神经网络不能识别自然语言文字,因此,本

文采用 Word2vec[25] 中 SKIP-Gram[26] 模型将分词结果转换为词向量。模型训练窗口设置为 5，词向量维度设置为 100，训练并行数设置为 2，迭代次数设置为 25。Word2vec 训练得到每个词的词向量储存到字典中，将水稻提问数据中的每个词转换成对应的词向量，例如“稻瘟病”的词向量如图 1 所示。由于每个问句的长度不是完全一致，将提问数据长度设置为 100，长度不足 100 的，在后面用 0 补齐，保证每一条提问数据向量长度一致，才能输入到神经网络中。文本预处理流程如图 2 所示。

稻瘟病

```
[ 6.7097824e-03  7.4307096e-01  9.0316826e-01  1.9036181e-01
 -4.3681994e-01 -7.0093483e-01 -1.2112090e-01 -6.3672201e-03
 -2.3016904e-03 -9.0564877e-02  5.7134330e-02  8.2240224e-01
 -1.9313449e-01 -5.5023742e-01 -5.1512951e-01  5.8883411e-01
 -7.5514615e-01 -3.1976292e-01 -3.4879699e-01 -3.3996466e-01
  2.1893804e-01 -7.2627313e-02 -1.1370174e-01  4.5059466e-01
 -6.7317314e-02 -5.6847054e-01 -6.0463142e-01  5.9052669e-02
 -1.3077798e-01 -1.6596270e-01  1.5853813e-01  6.5095556e-01
  6.4878941e-01  4.6016011e-01  7.5711823e-01  7.4668027e-02
 -1.8575861e-01 -7.3934114e-01  5.8915299e-01  1.2221171e+00
  1.1865857e+00  7.4959886e-01 -3.7220433e-02  5.1921362e-01
  3.1730467e-01  2.9835048e-01  1.9465286e-01  1.9594765e-01
  9.5289820e-01  8.8047618e-01 -1.8942514e-01 -1.9654460e-01
 -8.4836704e-01  6.5927351e-01  5.6682754e-01 -2.7341539e-01
 -3.4894038e-02 -2.9405308e-01 -2.1310528e-01 -1.1072905e+00
 -4.5399415e-01 -7.6834880e-02 -5.2894652e-01  7.3055312e-02
  7.4490570e-02 -5.8068454e-01  2.6303527e-01  2.6208603e-01
  8.2010448e-02  7.0894462e-01 -7.1187280e-02 -1.5448423e-01
  2.3968425e-01  2.4929245e-01 -1.5444218e-01  1.7619039e-01
  5.4907012e-01 -9.4590886e-01 -5.8262982e-01 -2.6189119e-01
  3.7055925e-02 -2.7945346e-01  4.1507450e-01 -2.7782950e-01
 -9.5942402e-01 -1.9637915e-01 -4.5042351e-01  3.6467716e-01
 -5.1423943e-01  4.3981120e-01 -2.1299433e-02 -9.5975882e-01
  9.5942760e-01 -3.9937019e-01  2.7054337e-01  7.6532632e-01
 -1.6501517e-04 -5.5885637e-01 -1.7992486e-01  8.8758342e-02]
```

图 1　“稻瘟病”词向量

Fig. 1　Word vector of rice blast

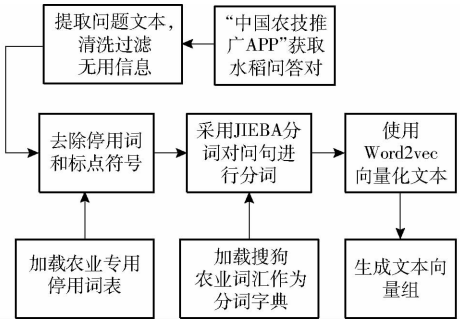

图 2　文本预处理流程图

Fig. 2　Flow chart of text preprocessing

1.4　基于注意力机制的密集连接卷积神经网络

本文提出的基于注意力机制的密集连接卷积模型(Attention_DenseCNN)，包括密集连接的卷积，注意力机制和 Softmax 分类层，模型结构如图 3 所示。首先将水稻问题文本数据预处理向量化之后输入到密集连接卷积网络层进行特征自动提取，密集连接 CNN 包含多个卷积块，每个卷积块都包含卷积、批次标准化(Batch normalization)和激活函数 ReLU，卷积块流程如图 4 所示。为了更进一步选取特征，将提取的特征输入到注意力机制层进行权重重置及特征选择，最后将提取的特征向量输入到分类层。

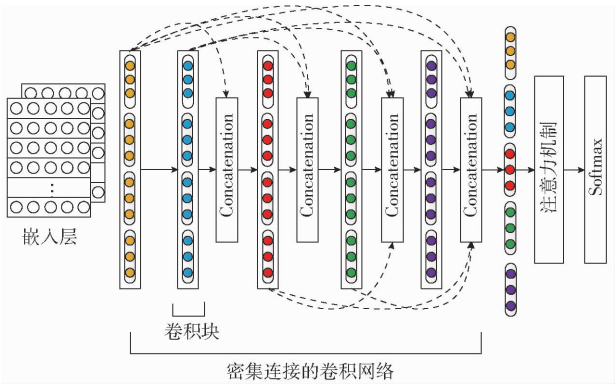

图 3　模型结构图

Fig. 3　Structure diagram of model

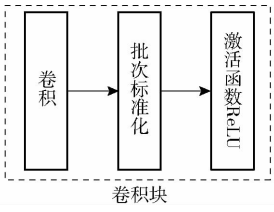

图 4　卷积块流程图

Fig. 4　Flow chart of convolution block

1.4.1　密集连接的 CNN

针对水稻提问文本数据特征的稀疏性，通过卷积块之间的密集连接加强文本特征的传输，通过特征的不断重复学习，增强文本特征在各卷积块之间的流动，使得模型能够自动提取和学习文本特征。密集连接的 CNN 允许网络的每一层直接连接到之前的所有层，以实现特征的重复使用，这种连接有助于提高水稻问句特征的整体利用率。同时将网络的每一层设计的非常狭窄，即只学习很少的特征映射来减少冗余。本文采用密集连接的 CNN 来提取问句的高层语义特征。每一层连接到通道维度中的上一层，并作为下一层的输入，每个层都会接收其前面所有层输出作为其额外的输入，传统的卷积神经网络模型都是采取固定的卷积核尺寸来提取特征，为了提取可变长度的文本特征，本文在每个卷积块都采用不同宽度的卷积核提取特征，卷积块包含卷积层、批量标准化和 ReLU 激活函数。卷积层主要提取水稻问句向量矩阵的特征。批量标准化可以防止梯度消失，提升学习收敛速度，缩短模型训练时间。ReLU 激活函数用来将输入的数据做非线性映射。每个卷积块可以形式化为

$$X_l = f(\boldsymbol{W}_l, [X_1, X_2, \cdots, X_{l-1}])$$
$$(\boldsymbol{W}_l \in \mathbf{R}^{(l-1)k^2w}) \qquad (1)$$

式中　f——复合函数，包括卷积、Batch normalization 和 ReLU 3 个步骤

l——卷积块层数　　X_l——第 l 层特征

$\boldsymbol{W}_l$——第 l 层的权重矩阵

k——卷积核个数　　w——窗口尺寸

比如已经得到 X_1 和 X_2，分别进行卷积操作得到 2 个 $m \times k$ 的矩阵，然后进行相加得到 X_3。除了输入，由于注意力机制的需要，所有特征图的尺寸都设置成 $m \times k$。这种结构就可以使得上游卷积块的词或者短语可以被后续结构使用，得到大尺度的特征。

1.4.2 注意力机制层

针对水稻提问文本数据长度较短，词汇量较少，文本稀疏性等特点，本文使用注意机制来有效地利用通过密集连接卷积神经网络得到的特征进行分类。在注意力机制层，利用注意力机制来强化关键词特征在水稻提问数据中的权重，解决了文本特征稀疏性导致的关键词权重较低的问题，注意力机制为每个特征向量分配不同的权重，以反映它们在问题中的作用。权重越大，语义越重要，并将影响整个问题的最终分类。该部分包含滤波器集合和尺度权重重置两步操作。连接起来之后经过 MLP 得到各部分的权重 $\boldsymbol{\alpha}_i$。具体计算公式为

$$\boldsymbol{s}^i = [s_1^i, s_2^i, \cdots, s_L^i] \tag{2}$$

$$\boldsymbol{\alpha}_L^i = \mathrm{MLP}(\boldsymbol{s}^i) \tag{3}$$

$$\boldsymbol{\alpha}_i = [\alpha_1^i, \alpha_2^i, \cdots, \alpha_L^i] \tag{4}$$

式中　$\boldsymbol{\alpha}_L^i$——第 L 层文本第 i 个位置对应的向量

MLP——多层感知器函数

s_L^i——第 L 层经过压缩之后的标量

$\boldsymbol{\alpha}_i$——用于输入分类器的权重向量

1.4.3 Softmax 分类层

在输出层，以注意力机制层的输出作为输入，并使用 Softmax 计算每种分类的概率，再将概率归一化到(0,1)，最后以最大概率输出水稻提问文本数据对应的具体类别。

1.5 模型训练

本文使用随机梯度下降(SGD)对模型参数进行优化。随机梯度下降算法是每次训练一个样本和类别标签对参数进行更新，其更新公式为

$$\varphi \Leftarrow \varphi - \eta \nabla_\varphi J(\varphi; x^{(i)}; y^{(i)}) \tag{5}$$

式中　$J(\varphi)$——目标函数

φ——参数　　η——学习速率

$x^{(i)}$——样本　　$y^{(i)}$——类别标签

$\nabla_\varphi J$——参数 φ 梯度

本文使用交叉熵损失函数来评判当前训练得到的概率分布与真实分布的差异性，如果该类别和样本的类别相同就是 1，否则是 0，交叉熵损失函数公式为

$$L(p, y) = -\sum_{c=1}^{M} y_c \lg p_c \tag{6}$$

式中　M——类别数量　　c——类别

y——指示变量，取 0 或 1

p——对于观测样本属于类别 c 的预测概率

2 试验及结果分析

2.1 硬件、软件环境和评价指标

试验软件环境为 Python 3.6.2 和 TensorFlow 1.13.1，服务器硬件环境为：显卡 NVIDIA Corporation Device 1e04 (rev a1)，GPU：NVIDIA GeForce RTX 2080Ti。使用 Keras 神经网络框架构建网络。从建立的水稻问答语料库中提取出水稻提问数据作为试验数据，将水稻提问文本数据进行人工标注。分为病虫害、草害药害、高产增产、储存保鲜、栽培管理、其他 6 个类别，共计 12 000 条提问数据，问题类别数据具体分布见表 3。将 12 000 条提问数据随机打乱顺序后，按照 9∶1 的比例划分为训练集和测试集，训练集共 10 800 条，测试集 1 200 条，训练集用于模型训练及调整神经网络参数，测试集用来验证模型效果及反馈时间。本文以精确率 P、召回率 R 和 F1 值作为评价指标。

表 3 水稻问题类别分布

Tab.3 Distribution of rice problems

问题类别	病虫害	草害药害	高产增产	储存保鲜	栽培管理	其他
数量/条	4 312	1 398	1 350	980	1 983	1 977

2.2 参数设置

通过 Word2vec 转换的词向量维度设置为 100，问句最大长度设置为 100，神经网络的输入就是(100×100)维的向量矩阵，模型训练次数设置为 50 轮，每批次输入 64 条，学习速率设置为 0.01。

由 1.4 节可知，密集连接的卷积神经网络拥有多个卷积块，通过一组试验来测试在不同数量卷积块下模型的分类效果，设置卷积块个数分别为 3、4、5、6，精确率分别为 94.3%、94.7%、95.0%、94.6%，可知当卷积块个数为 5 时，模型分类效果最佳。

卷积模型在提取文本特征时，卷积核的长度与文本向量的维度相同为 100，在不同卷积块中采用不同的卷积核宽度来提取可变的文本特征。通过一组试验来研究不同宽度的卷积核对问句分类精确率的影响。如表 4 所示，在 5 个卷积块中，卷积核宽度分别为 2、3、4、5、6 时，模型的分类精确率最高。

通过一组试验来研究卷积滑动窗的步长对 Attention_DenseCNN 模型影响。设置步幅分别为 1、2、3、4。除了步幅，其他参数都保持不变。本文所提

表 4　不同卷积核宽度下模型分类效果

Tab. 4　Classification effect of models with different convolution kernel widths

卷积核宽度	精确率/%
1、2、3、4、5	93.6
2、3、4、5、6	95.5
1、3、4、5、6	94.7
1、2、3、4、6	94.9
1、2、4、5、6	93.9

出的模型在卷积步幅分别为 1、2、3、4 时的问句分类精确率分别为 95.0%、95.6%、94.2%、92.7%。在步幅增大过程中，Attention_DenseCNN 对于数据集的分类精确率会在步幅为 2 时达到 95.6%。相比于步幅为 1 时精确率提高了 0.6 个百分点，这是由于卷积滑动窗口在步幅为 2 时能更好地提取文本特征，表达文本信息。而在步幅为 3 和 4 时，精确率开始下降，分别为 94.2% 和 92.7%。这是由于步幅过大，在提取特征时会丢失更多的语义信息。

2.3　问句分类模型分类结果与评价

首先使用密集连接卷积神经网络（DenseCNN）、基于文本的卷积神经网络（TextCNN）[15]、长短期记忆网络（LSTM）[27]、门控循环单元网络（BiGRU）[28]、基于注意力机制的长短期记忆网络（Attention－LSTM）[29] 5 种神经网络模型对水稻问句数据进行试验。表 5 展示了不同深度学习模型在精确率、召回率和 F1 值的比较。从表 5 可以看出，在基本的对比试验中，TextCNN 与 3 个循环神经网络的 F1 值相似，证明了卷积神经网络在水稻问题分类中有很好的性能。DenseCNN 获得了最高的 F1 值和精确率，这是由于通过密集连接的卷积块可以加强特征的传递和提取，减少特征损失，有利于最后的分类效果。基于注意力机制的 LSTM 相比于 LSTM 有更好的精确率、召回率和 F1 值，但略低于 DenseCNN，这说明了注意力机制在训练过程中通过权重重置，可以更好地表达特征信息。

表 5　5 种神经网络模型效果对比

Tab. 5　Comparison of five neural network models

%

模型	精确率	召回率	F1 值
TextCNN	92.7	89.6	91.1
LSTM	89.3	87.7	88.5
DenseCNN	93.7	92.4	93.0
BiGRU	91.6	88.6	90.1
Attention－LSTM	90.3	89.9	90.0

将注意力机制与上文提到的 4 种神经网络模型相结合，得到 Attention_DenseCNN 模型、Attention－TextCNN 模型[30]、Attention－LSTM 模型[29]、Attention－BiGRU[31] 模型，进行水稻问题分类试验。表 6 展示了不同深度学习模型在精确率、召回率和 F1 值的比较。通过比较每种模型的 F1 值及精确率，可以发现在水稻问句文本数据集下，DenseCNN 模型在训练效果上优于 TextCNN 模型，这是由于在 DenseCNN 模型中是通过密集连接的 5 个卷积块来提取特征。相比于传统的卷积神经网络模型在特征传递时只是将上一层的输出作为本层的输入，DenseCNN 在传递特征时不仅接收上一层的输出，还将接收之前所有层的输出作为额外输入，可以有效减少文本特征的丢失。通过密集连接的 CNN，文本特征可以更好地传递和表达，最终提高文本分类效果。从表 6 可以看出，Attention_DenseCNN 模型的精确率、召回率和 F1 值最高。其精确率为 95.6%，召回率为 94.3%，F1 值为 94.9%。F1 值比 Attention－TextCNN 提高 2.8 个百分点。将加入注意力机制后的 4 种模型与未加入前的模型进行对比分析，可以发现，精确率分别提升了 0.9、1.9、1、0.7 个百分点，F1 值分别提升了 1、1.9、1.5、1.3 个百分点。这说明注意力机制能使模型在训练过程高度关注目标的特征信息，从而更好地识别特定目标的类别，验证了注意力机制在问句分类任务中的有效性。

表 6　不同模型下问句分类效果对比

Tab. 6　Comparison of question classification between different models

%

模型	精确率	召回率	F1 值
TextCNN	92.7	89.6	91.1
Attention－TextCNN	93.6	90.6	92.1
DenseCNN	93.7	92.4	93.0
Attention_DenseCNN	95.6	94.3	94.9
LSTM	89.3	87.7	88.5
Attention－LSTM	90.3	89.9	90.0
BiGRU	91.6	88.6	90.1
Attention－BiGRU	92.3	90.6	91.4

图 5 和图 6 分别展示了 8 种神经网络模型的 F1 值趋势。Attention_DenseCNN 的收敛时间明显提前，并且每个模型在训练过程中不会出现大的振荡。另外，Attention_DenseCNN 在前 20 个轮次大致可以收敛，在随后的训练中 F1 值保持最高。

表 7 展示了使用 Attention_DenseCNN 模型对水稻提问文本数据，在草害药害、高产增产、病虫害、储存保鲜、栽培管理、其他共 6 个类别分类的精确率、召回率及 F1 值。Attention_DenseCNN 模型分类平均精确率达到了 95.6%，召回率达到了 94.3%，F1 值达到了 94.9%，体现了良好的分类效果。本文提

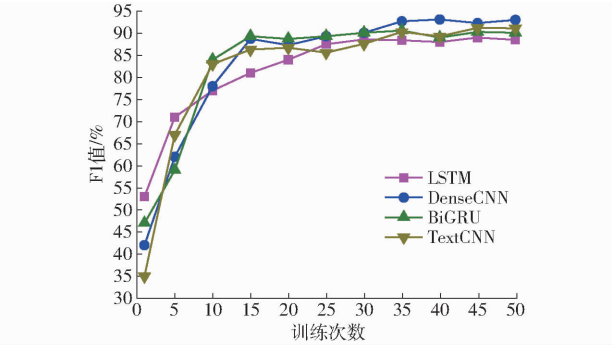

图 5 4 种基本神经网络 F1 值趋势图

Fig. 5 F1 value trend graph of four basic neural networks

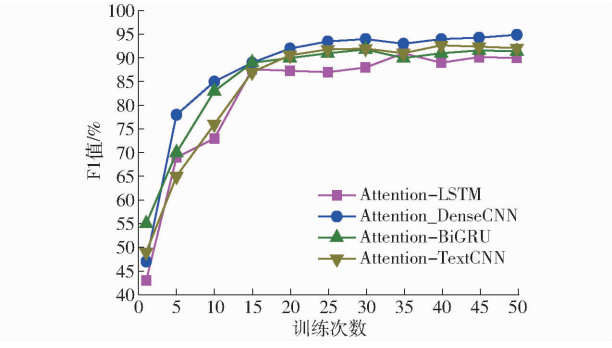

图 6 4 种基于注意力机制的神经网络模型 F1 值趋势图

Fig. 6 F1 value trend graph of four attention mechanism of neural network models

表 7 Attention_DenseCNN 模型分类结果

Tab. 7 Classification results of Attention_DenseCNN model %

类别	精确率	召回率	F1 值
草害药害	97.1	97.3	97.2
高产增产	95.6	95.8	95.7
病虫害	99.1	99.0	99.0
储存保鲜	90.2	89.6	89.9
栽培管理	97.1	92.6	94.8
其他	94.5	91.8	93.1

出模型在病虫害类别中精确率和 F1 值明显高于其他 5 个类别,达到了 99.1% 和 99%,这是由于深度学习模型需要大量的数据支撑,病虫害作为水稻种植过程中经常遇到的问题,其文本数据占到整体文本数据的 1/3,因此分类精确率及 F1 值最高,而在储存保鲜类别中精确率及 F1 值最低,说明数据量过小,会影响模型的分类效果。

表 8 展示了基于注意力机制的 4 种神经网络模型在 1 200 条测试集上的响应时间和精确率,达到了对水稻问句快速分类的要求,Attention - TextCNN 的响应时间最短,这是由于 Attention - TextCNN 模型结构简单,训练层数和模型参数较少。本文提出的 Attention_DenseCNN 模型在 5 s 可以完成对测试集 1 200 条水稻问句进行精准分类,精确率达到了最高(94.7%),在响应时间相差不大的情况下,本文提出的模型在精确率方面具有更好的效果。

表 8 4 种网络模型的响应时间和精确率

Tab. 8 Response time and accuracy of four network models

模型	响应时间/s	精确率/%
Attention - TextCNN	3	91.8
Attention - LSTM	4	88.7
Attention - BiGRU	4	89.9
Attention_DenseCNN	5	94.7

3 结论

(1)采用 Word2vec 和农业词汇专业词典对文本进行向量化,有效解决了水稻问句文本的稀疏性和高维性。使用基于注意力机制的密集连接卷积神经网络构建水稻问答系统问句自动分类模型,可以快速、准确地实现水稻问句在 6 个类别上的自动分类,避免了浅层学习模型耗时、分类效果不佳的问题,模型精确率达到 95.6%,F1 值达到 94.9%,与其他 7 类文本分类模型相比优势明显。

(2)使用注意力机制对文本关键词特征进行权重重置,强化了关键词在水稻问句中的作用,切实解决了水稻问句特征不足的问题,提出的 Attention_DenseCNN 相比 DenseCNN,水稻问句文本分类精确率和 F1 值均提高了 1.9 个百分点。

参 考 文 献

[1] 奚雪峰,周国栋. 面向自然语言处理的深度学习研究[J]. 自动化学报, 2016,42(10):1445 - 1465.
XI Xuefeng, ZHOU Guodong. A survey on deep learning for natural language processing[J]. Acta Automatica Sinica, 2016, 42(10):1445 - 1465. (in Chinese)

[2] WASIM M, ASIM M N, GHANI KHAN M U, et al. Multi-label biomedical question classification for lexical answer type prediction[J]. Journal of Biomedical Informatics, 2019,93:103 - 143.

[3] FIGUEROA A, NEUMANN G. Context-Aware semantic classification of search queries for browsing community question-answering archives[J]. Knowledge-Based Systems,2016,96(15):1 - 13.

[4] AHMED W, ANTO B. Answer extraction technique for question answering systems[J]. International Journal of Innovative Research in Computer and Communication Engineering,2016,4(11):20352 - 20357.

[5] SARROUTI M, SAID O. SemBioNLQA: a semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions[J]. Artificial Intelligence in Medicine, 2020, 102:101767.1 - 101767.16.

[6] 李道亮,杨昊. 农业物联网技术研究进展与发展趋势分析[J/OL]. 农业机械学报,2018,49(1):1 - 20.
LI Daoliang, YANG Hao. State-of-the-art review for internet of things in agriculture[J/OL]. Transactions of the Chinese Society for Agricultural Machinery,2018,49(1):1 - 20. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? flag=1&file_

no = 20180101&journal_id = jcsam. DOI:10. 6041/j. issn. 1000-1298. 2018. 01. 001. (in Chinese)
[7] 段青玲,张璐,刘怡然,等. 基于农业网络信息分类的热词自动提取方法[J/OL]. 农业机械学报,2018,49(7):160 - 167.
DUAN Qingling, ZHANG Lu, LIU Yiran, et al. Automatic extraction method of hot words based on agricultural network information classification[J/OL]. Transactions of the Chinese Society for Agricultural Machinery,2018,49(7):160 - 167. http://www. j-csam. org/jcsam/ch/reader/view_abstract. aspx? flag = 1&file_no = 20180720&journal_id = jcsam. DOI: 10. 6041/j. issn. 1000-1298. 2018. 07. 020. (in Chinese)
[8] LIU P, ZHAO H H, TENG J Y, et al. Parallel naive Bayes algorithm for large-scale Chinese text classification based on spark [J]. Journal of Central South University, 2019, 26(1):1 - 12.
[9] PANDEY C, IBRAHIM Z, WU H, et al. Improving RNN with attention and embedding for adverse drug reactions[C]// Proceedings of the 2017 International Conference on Digital Heath,2017: 67 - 71.
[10] 陈瑛,陈昂轩,董玉博,等. 基于 LSTM 的食品安全自动问答系统方法研究[J/OL]. 农业机械学报,2019,50(增刊):380 - 384.
CHEN Ying,CHEN Angxuan,DONG Yubo,et al. Methods of food safety question answering system based on LSTM[J/OL]. Transactions of the Chinese Society for Agricultural Machinery,2019,50(Supp.):380 - 384. http://www. j-csam. org/jcsam/ch/reader/view_abstract. aspx? flag = 1&file_no = 2019s058&journal_id = jcsam. DOI:10. 6041/j. issn. 1000-1298. 2019. S0. 058. (in Chinese)
[11] 梁敬东,崔丙剑,姜海燕,等. 基于 word2vec 和 LSTM 的句子相似度计算及其在水稻 FAQ 问答系统中的应用[J]. 南京农业大学学报,2018,41(5): 946 - 953.
LIANG Jingdong,CUI Bingjian,JIANG Haiyan, et al. Sentence similarity computing based on word2vec and LSTM and its application in rice FAQ question-answering system[J]. Journal of Nanjing Agricultural University, 2018, 41(5): 946 - 953. (in Chinese)
[12] ZHOU C, SUN C, LIU Z, et al. AC - LSTM neural network for text classification [J]. Computer Science, 2015, 1(4):39 - 44.
[13] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification[C]// Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics,2016.
[14] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv Preprint arXiv: 1408. 5882, 2014.
[15] LIU P, QIU X, HUANG X. Recurrent neural network for text classification with multi-task learning[J]. arXiv Preprint arXiv: 1605. 05101, 2016.
[16] HUANG G, LIU Z, LAURENS V D M, et al. Densely connected convolutional networks[J]. arXiv Preprint arXiv:1608. 06993, 2016.
[17] SONG J, HUANG X, QIN S, et al. A bi-directional sampling based on K-means method for imbalance text classification[C]// Proceedings of 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS),2016:1 - 5.
[18] FENG G, GUO J, JING B Y, et al. Feature subset selection using naive Bayes for text classification[J]. Pattern Recognition Letters, 2015, 65(1):109 - 115.
[19] HADDOUD M, MOKHTARI A, LECROP T, et al. Combining supervised term-weighting metrics for SVM text classification with extended term representation[J]. Knowledge & Information Systems, 2016, 49(3):1 - 23.
[20] 赵明,董翠翠,董乔雪,等. 基于 BIGRU 的番茄病虫害问答系统问句分类研究[J/OL]. 农业机械学报,2018,49(5):271 - 276.
ZHAO Ming,DONG Cuicui,DONG Qiaoxue,et al. Question classification of tomato pests and diseases question answering system based on BIGRU[J/OL]. Transactions of the Chinese Society for Agricultural Machinery,2018,49(5):271 - 276. http://www. j-csam. org/jcsam/ch/reader/view_abstract. aspx? flag = 1&file_no = 20180532&journal_id = jcsam. DOI:10. 6041/j. issn. 1000-1298. 2018. 05. 032. (in Chinese)
[21] 魏芳芳, 段青玲, 肖晓琰,等. 基于支持向量机的中文农业文本分类技术研究[J/OL]. 农业机械学报, 2015,46(增刊):174 - 179.
WEI Fangfang,DUAN Qingling,XIAO Xiaoyan,et al. Classification technique of Chinese agricultural text information based on SVM [J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2015, 46(Supp.):174 - 179. http://www. j-csam. org/jcsam/ch/reader/view_abstract. aspx? flag = 1&file_no = 2015S029&journal_id = jcsam DOI: 10. 6041/j. issn. 1000-1298. 2015. S0. 029. (in Chinese)
[22] 张明岳,吴华瑞,朱华吉. 基于卷积模型的农业问答语性特征抽取分析[J/OL]. 农业机械学报,2018,49(12):203 - 210.
ZHANG Mingyue,WU Huarui,ZHU Huaji. Analysis of extraction of semantic feature in agricultural question and answer based on convolutional model [J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2018,49(12):203 - 210. http://www. j-csam. org/jcsam/ch/reader/view_abstract. aspx? file_no = 20181226&flag = 1. DOI: 10. 6041/j. issn. 1000-1298. 2018. 12. 026. (in Chinese)
[23] 金宁,赵春江,吴华瑞,等. 基于 BiGRU_MulCNN 的农业问答问句分类技术研究[J/OL]. 农业机械学报,2020,51(5):199 - 206.
JIN Ning,ZHAO Chunjiang,WU Huarui, et al. Classification technology of agricultural question based on BiGRU_MulCNN[J/OL]. Transactions of the Chinese Society for Agricultural Machinery,2020,51(5):199 - 206. http://www. j-csam. org/jcsam/ch/reader/view_abstract. aspx? flag = 1&file_no = 20200522&journal_id = jcsam. DOI:10. 6041/j. issn. 1000-1298. 2020. 05. 022. (in Chinese)
[24] 赵静. 大规模汉语语义词典构建[D]. 哈尔滨:哈尔滨工业大学,2011.
ZHAO Jing. Building a large scale Chinese semantic dictionary [D]. Harbin: Harbin Institute of Technology, 2011. (in Chinese)
[25] GOLDBERG Y, LEVY O, MIKOLOV,et al. Word2vec explained: deriving negative-sampling word-embedding method [J]. arXiv Preprint arXiv: 1402. 3722,2014.
[26] YONGWHI K A, GWAN H J, HODUK C,et al. Automated text analysis based on Skip-Gram model for food evaluation in predicting consumer acceptance[J/OL]. Computational Intelligence and Neuroence, 2018:9293437. https://doi. org/10. 1155/2018/9293437.
[27] LI Pengfei, QIU Xipeng, CHEN Jifan, et al. Deep fusion lstms for text semantic matching[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,2016:1034 - 1043.
[28] TANG Q, LI J, CHEN J, et al. Full attention-based Bi-GRU neural network for news text classification[C]//2019 IEEE 5th International Conference on Computer and Communications (ICCC). IEEE, 2019.
[29] BAZIOTIS C, PELEKIS N, DOULKERIDIS C. DataStories at semEval - 2017 task 6: siameselstm with attention for humorous text comparison[C]// Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval - 2017), 2017:747 - 754.
[30] HE T, HUANG W, QIAO Y, et al. Text-attentional convolutional neural network for scene text detection[J]. IEEE Trans. Image Process,2016,25:2529 - 2541.
[31] LIU J, YANG Y, LV S,et al. Attention-based Bi-GRUCNN for Chinese question classification[J/OL]. Journal of Ambient Intelligence and Humanized Computing, 2019. https://doi. org/10. 1007/s12652 - 019 - 01344 - 9.