Advances in Applied Mathematics 应用数学进展, 2022, 11(10), 7348-7361
Published Online October 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.1110781

Hans汉斯

深度学习算法在新闻文本分类中的应用

李　伟*，王丙硕，林明志

西安电子科技大学，数学与统计学院，陕西　西安

收稿日期：2022年9月21日；录用日期：2022年10月14日；发布日期：2022年10月25日

摘　要

新闻文本分类是将新闻文本划分为不同类别的多分类任务，旨在识别文本的关键语义信息，帮助用户获取目标新闻。基于深度学习方法，本文构建了一种新的新闻文本分类模型textPDCNN，该模型主要包括LSTM层、多尺度空洞卷积层、池化层和全连接层。其中，LSTM层主要用于提取文本的语义信息，多尺度空洞卷积层用来融合不同膨胀率的空洞卷积所获取的不同尺度信息，池化层用于提取最有利于分类的特征，全连接层将提取到的特征映射到分类空间。实验表明：相比传统的HAN (Hierarchical Attention Networks)模型，textPDCNN文本分类模型的*Macro_P*，*Macro_R*和*Macro_F*指标分别提升了0.74%、0.61%和0.49%。并且，textPDCNN在公共数据集THUCNews上的性能超过目前已知文献中的最好结果。总之，该方法能很好地融合不同长度的短语特征，并在新闻文本分类任务上具有显著优势。

关键词

深度学习，文本分类，空洞卷积，空洞空间金字塔池化

Application of Deep Learning Agorithm in News Text Classification

Wei Li*, Bingshuo Wang, Mingzhi Lin

School of Mathematics and Statistics, Xidian University, Xi'an Shaanxi

Received: Sep. 21st, 2022; accepted: Oct. 14th, 2022; published: Oct. 25th, 2022

Abstract

News text classification is a multi-classification task that divides news text into different categories. It aims to identify the key semantic information of the text and provide users with convenient access to target news. Based on the theory of deep learning, this paper constructs a news text clas-

*通讯作者。

文章引用: 李伟, 王丙硕, 林明志. 深度学习算法在新闻文本分类中的应用[J]. 应用数学进展, 2022, 11(10): 7348-7361.
DOI: 10.12677/aam.2022.1110781

sification model called textPDCNN, which mainly includes LSTM layer, multi-scale hole convolution layer, pooling layer and fully connected layer. LSTM layer is mainly used to extract the semantic information of the text, the multi-scale hole convolution layer mainly is used for integrating the information of different scales obtained by the hole convolution with different expansion rates. The pooling layer is used to extract features that are most conducive to classification, and the fully connected layer is to map the extracted features to the classification space. Experiments show that compared with traditional HAN model, the *Macro_P*, *Macro_R* and *Macro_F* indicators of textPDCNN have increased by 0.74%, 0.61% and 0.49% respectively. The performance of textPDCNN on the public data set THUCNews exceeds the best results currently known in the literature. To sum up, the model we proposed can integrate phrase features of different lengths well, and has obvious advantages in news text classification tasks.

Keywords

Deep Learning, Text Classification, Atrous Convolution, Atrous Spatial Pyramid Pooling

1. 引言

互联网和人工智能技术的蓬勃发展，给人们的生活增添异彩。比如自动驾驶技术、扫地机器人、智能推荐、自动分类等。并且人们逐渐在网络上进行信息交流和信息获取，因此网络上的信息数量呈爆炸式的增长。这些数量庞大的信息主要以视频、图片和文本的形式存在，其中以文本形式存在的信息占据非常大的比重。但是数量庞大的信息给人们获取目标信息带来巨大的困扰。因此，如何让用户能在杂乱无章、数量繁多的信息中快速获取想要的信息，成为广大研究者关注的课题。

文本分类的主要任务就是按照一定的规则将文本信息划分为不同的类别。对信息进行分类整理，使用户能够快速获取目标信息，为用户提供方便，减少了用户浏览垃圾信息的时间，从而降低了用户的时间成本。目前文本分类具有非常广泛的应用，比较典型的包括垃圾邮件分类、情感分析、信息检索和新闻文本分类等[1]。垃圾邮件分类系统[2]主要通过文本分类技术来判定该邮件是否为垃圾邮件。如果判定该邮件为垃圾邮件，那么系统会自动帮用户过滤掉该邮件，如果判定该邮件不是垃圾邮件，那么系统则会将邮件发送给收件人，从而节省用户阅读垃圾邮件的时间，提高用户的使用体验。情感分析[3]是指对具有感情偏向的文本根据其情感偏向程度划分为不同的类别，比如可将商品评论分为好评和差评。信息检索技术[1]是先将数据库中文本划分好类别，然后根据搜索内容所属类别去相应的类别库中搜索，这样可以通过缩小搜索范围，从而提高搜索速度。新闻文本分类[4]是根据新闻的文本内容将新闻划分为不同的主题，比如可以把新闻文本分为财经、家居、房产、科技和教育等，那么用户在阅读浏览新闻时可选择特定的类别进行浏览，从而节省用户的浏览时间。

根据中国互联网络信息中心发布的报告显示，我国网络新闻用户数量增长迅速，目前用户规模已高达 7.43 亿，占总体网民的 75.1%。中国网络新闻用户数量持续增加，新闻软件层出不穷，大量的纸质媒体也开始拓展网络新闻业务，比如国务院客户端、人民日报客户端、中国新闻网客户端和新华社客户端等。过去的纸质媒体会有专门的板块进行不同类别的新闻报道，现在的客户端会在指引页将不同类别的新闻区分开来。用户可以通过点击按钮来选择浏览不同类别的新闻。比如人民日报客户端将新闻分为财

经、体育、文化、教育、军事和科技等板块。过去主要是新闻编辑阅读浏览新闻后将新闻进行分类，目前新闻数量呈指数增长，如果仍使用人工的方法进行分类，会耗费大量的人力物力资源。如果使用文本分类技术，并能将收集的文本快速地进行分类，时间和人力物力资源将会大大节省。

文本分类被广泛地应用在人们生活的方方面面，因此文本分类是自然语言处理领域非常具有研究价值的方向之一。最开始人们依靠专家设定的判别规则来进行文本分类[5]，到后来人们根据人工提取文本特征的方法即传统机器学习方法来进行文本分类[6]。但是这些方法基本都依靠专家知识来设定规则和提取特征，因此其性能主要取决于专家知识的好坏。并且这些方法无法提取到词与词之间的深层次特征，导致出现泛化能力不强、鲁棒性较差的问题。当前，深度学习在人工智能领域取得了巨大的进展。通过神经网络非凡的特征提取能力，深度学习模型可以提取到信息深层次的特征，从而解决了许多传统方法无法解决的问题。比如深度学习可以解决文本分类的语义表征和特征稀疏等问题。深度学习被应用到文本分类领域，使得文本分类的准确率比传统方法上升了一个维度。

深度学习模型的效果主要取决于所用的神经网络结构，本文将构建适用于提取新闻文本特征的神经网络结构来进一步提高新闻文本分类的准确率。本文的工作主要如下：

1) 提出一维空洞卷积概念，将其用于新闻文本分类模型中，并通过提高模型感受野来提高模型效果。

2) 提出基于空洞空间金字塔池化的新闻文本分类模型，该模型首先通过 LSTM 提取文本语义，然后使用不同膨胀率的空洞卷积来提取文本不同长度的短语特征，最后使用池化提取文本的显著特征用于文本分类。

3) 使用清华大学开源中文文本分类数据集 THUCNews 对模型进行实验验证。

2. 相关工作

文本分类本质上就是通过一定的方法来获取数据集到其类别标签的映射关系或模型，并且可以通过该映射关系或模型实现对未知类别标签的自动分类[7]。文本分类的研究主要可以分为三个阶段[8]。

第一阶段为 20 世纪 60 年代到 20 世纪 80 年代，该阶段主要通过人工对数据集的分析来确定每个类别的分类规则；比如 if-then 规则，通过对不同类别的文本设定不同的规则来实现文本与类别的匹配，从而确定文本的类别标签。该类型的方法主要包括决策树、关联规则等[8]。决策树分类器[9]主要通过对数据空间的层次分解来创建一个基于文本属性的树来进行分类。赵琳学者[10]通过挖掘有利于分类的关联规则，来对规则进行组合，从而提高分类的效果。基于规则的文本分类方法虽然无需训练成本，但模型的性能主要取决于专家的知识，并且适用性较差，一般只适用于某一专门领域内的文本分类，相对来说泛化能力较弱。

第二阶段是 20 世纪 90 年代到 21 世纪初，研究学者们不再通过简单的规则匹配来进行文本分类，而是通过分析训练样本特征与类别标签之间的关系，来对测试样本进行标签的预测。这些方法主要包括支持向量机[11]、朴素贝叶斯[12]、K 近邻[13]等。这些基于统计的机器学习方法，需要人工构建繁杂而又低效的特征工程，这个过程需要消耗大量的人力物力资源，并且这些方法本质上还是人类知识驱动的方法，模型的性能主要依赖于人工提取的特征，对于简单的文本分类可以达到不错的效果，但对复杂的文本特征表达能力有限，导致模型泛化能力不强且鲁棒性较差。

第三阶段为 21 世纪初至今，深度学习在文本分类领域取得不错的成绩。深度学习的方法去掉了复杂繁琐的人工提取特征的步骤，转而使用神经网络去提取文本的深层语义特征，从而有利于提高模型的效率和准确率。常用的神经网络结构包括循环神经网络(RNN)、卷积神经网络(CNN)和注意力机制(Attention)等[14]循环神经网络将序列前面的特征计算后作为输入传递到后面，从而使网络能够捕捉到完整的序列上下文信息，有利于提取文本的完整语义信息，从而有利于提高模型准确率；比如 2016 年 Liu 等人[15]提

出了基于 RNN 的文本分类模型 textRNN，textRNN 主要是使用双向 LSTM 来提取全文的语义，该模型在文本分类上效果不错。卷积神经网络之前被广泛的运用在图像处理领域，2014 年，Kim [16]提出 textCNN 模型，将一维卷积运用到文本分类领域，并且取得了不错的效果。由于注意力机制其强大的特征提取能力，也被广泛的应用在文本分类领域，Yang 等人[17] 2016 年提出分层注意力模型 HAN，分层提取文档的词语和句子语义，在许多数据集上都取得了良好的性能。

3. 基于空洞空间金字塔池化的新闻文本分类模型

本文提出的基于空洞空间金字塔池化的新闻文本分类模型(textPDCNN)是用于解决新闻的长文本分类问题。该模型结构主要包括输入层、LSTM 层、多尺度空洞卷积层、池化层和全连接层。输入层是对新闻文本的文本表示层，LSTM 层主要用于提取新闻文本的语义关系，多尺度空洞卷积层是使用不同尺寸的卷积来提取文本不同长度的短语特征，并且本文使用一维空洞卷积来提高模型的感受野，从而提取更长的短语之间的依赖关系，池化层主要用于提取文本的关键特征用于文本分类。模型的结构如图 1 所示：

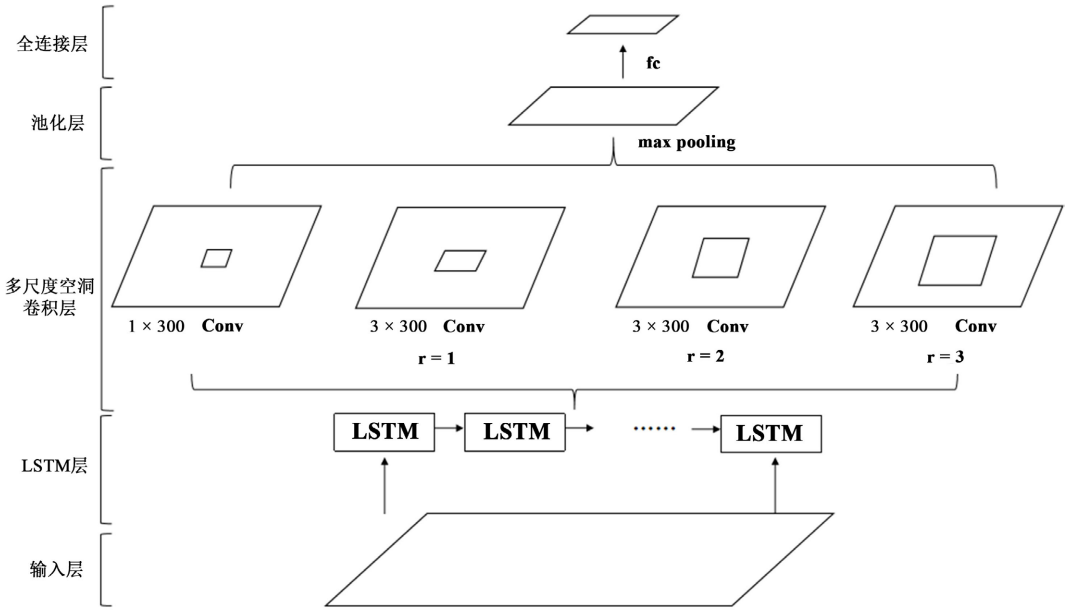

Figure 1. textPDCNN model structure
图 1. textPDCNN 模型结构图

输入层是使用预训练词向量对样本的文本表示，其中预训练词向量的维度为 300。本文选择将文本长度固定为 600，文本预处理会将文本长度大于 600 的样本进行裁剪操作，会将文本长度小于 600 的样本进行填充操作。本文选择的预训练词向量的维度为 300，故输入层的维度为[600, 300]。令 x_i 为词向量，那么输入矩阵可以表示为：

$$X = \left[x_1, x_2, \cdots, x_{600}\right]^{\mathrm{T}} \tag{1}$$

LSTM 通过门控开关来控制存储在细胞状态中的信息，因而 LSTM 可以获取文本中的远距离依赖关系；此外 LSTM 当前时刻的输出不仅与当前时刻的输入有关，而且与前一时刻的隐藏状态有关，因而 LSTM 可以获取文本的完整语义信息。LSTM 层的计算公式为：

$$\begin{aligned}
\tilde{c}_t &= \tanh\left(W\cdot\left[x_t,h_{t-1}\right]\right)\\
f_t &= \sigma\left(W^f\cdot\left[x_t,h_{t-1}\right]\right)\\
i_t &= \sigma\left(W^i\cdot\left[x_t,h_{t-1}\right]\right)\\
o_t &= \sigma\left(W^o\cdot\left[x_t,h_{t-1}\right]\right)\\
c_t &= \left(f_t\otimes c_{t-1}\right)\oplus\left(i_t\otimes\tilde{c}_t\right)\\
h_t &= o_t\otimes\tanh\left(c_t\right)
\end{aligned} \tag{2}$$

其中，c_t、f_t、i_t、o_t 分别指记忆门、遗忘门、输入门和输出门，W、W^f、W^i、W^o 指权重矩阵，x_t 指该时刻的输入，h_{t-1} 指前一时刻的输出，$\left[x_t,h_{t-1}\right]$ 指两个矩阵进行拼接，σ 代表 sigmoid 函数，c_{t-1} 指前一时刻的细胞状态，h_t 指该时刻的输出，$\oplus$ 指向量相加，$\otimes$ 指向量相乘。sigmoid 函数将门控的输出转换为 0 到 1 之间的数值；f_t 代表遗忘门，主要决定遗忘上一个细胞状态 c_{t-1} 的哪些信息；i_t 代表输入门，主要决定将哪些信息储存在新的细胞状态中，所以新的细胞状态 c_t 中包括两部分信息，一部分为上个细胞状态保留的信息，一部分为选择要储存的输入信息；o_t 代表输出门，主要决定在细胞状态中选择哪些信息进行输出。LSTM 层的计算公式如下所示：

$$\begin{cases}
h_i = \mathrm{LSTM}\left(x_i\right)\\
h = \left[h_1,h_2,\cdots,h_{600}\right]^{\mathrm{T}}
\end{cases} \tag{3}$$

研究学者提出一维卷积神经网络，用于自然语言处理领域词向量的特征提取，并且取得了不错的进展。空洞卷积目前广泛的应用于图像处理的语义分割领域，并使该领域模型效果得到不错的提升。因此本文提出一维空洞卷积，并将其使用在新闻文本分类领域。一维空洞卷积是结合一维卷积和空洞卷积的特点，一维卷积卷积核的宽等于输入矩阵的宽，因而一维卷积只会上下移动，不会左右移动；空洞卷积是在普通卷积核中注入空洞，从而增加感受野。因此本文提出的一维空洞卷积，卷积核的宽等于输入矩阵的宽，并且在注入空洞时，只会在行向量方向上注入空洞，而不会在列向量方向注入空洞，即空洞卷积的膨胀率为(r, 1)，代表在横向量方向的膨胀率为 r，在列向量方向的膨胀率为 1，本文提到的膨胀率为横向量方向的膨胀率，默认列向量方向的膨胀率为 1。膨胀率分别为 1、2、3 的一维空洞卷积示意图如图 2 所示：

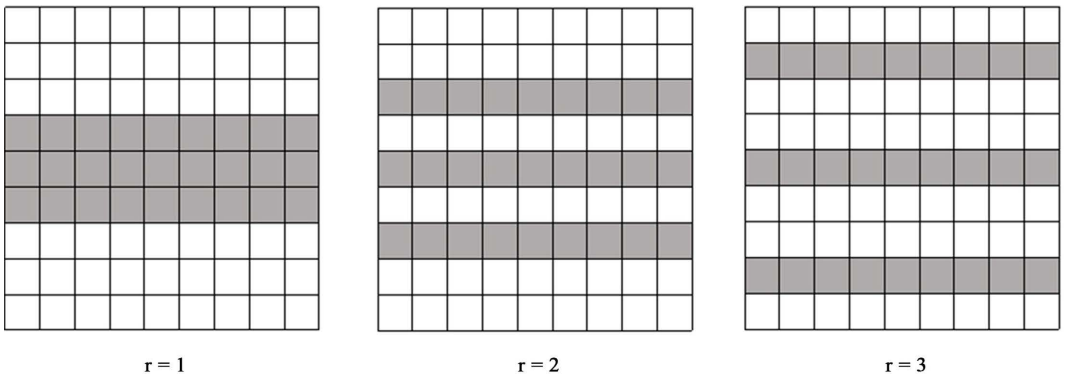

Figure 2. Sketch of one-dimensional cavity convolution
图 2. 一维空洞卷积示意图

在新闻文本分类领域，文本的长度往往非常长，而且语言之间的联系非常远，因而使用卷积神经网络往往不能捕捉到文本的远距离依赖关系，从而导致卷积神经网络模型的正确率不高。为解决卷积神经

网络感受野受限的问题，在不增加卷积神经网络参数的情况下，本文将一维空洞卷积神经网络应用到新闻文本分类领域。分别使用不同尺寸的空洞卷积对 LSTM 的输出进行操作，多尺度空洞卷积层示意图如图 3 所示：

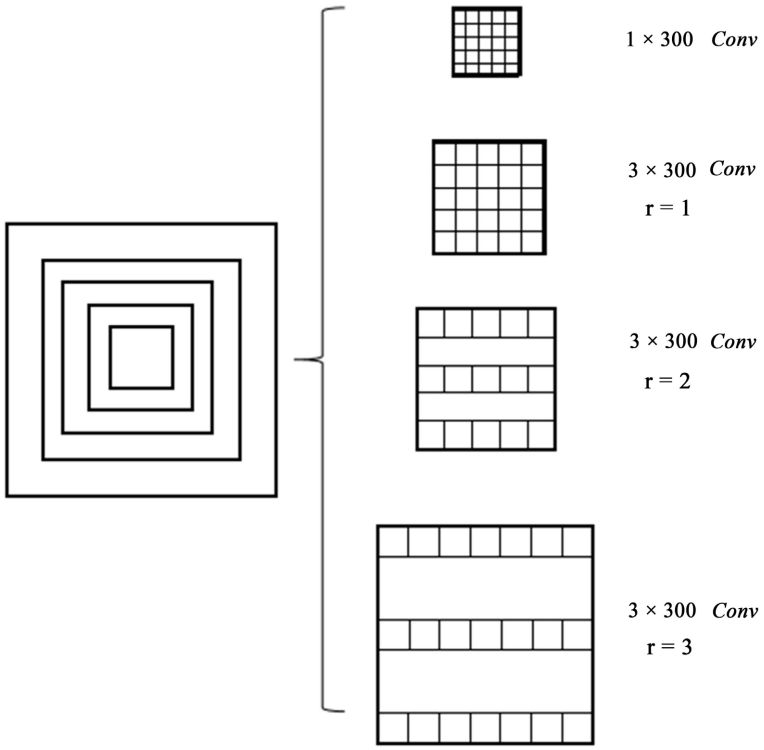

Figure 3. Sketch of multi-scale cavity convolution
图 3. 多尺度空洞卷积示意图

空洞卷积层的计算公式如下所示：

$$\begin{aligned} C_i &= \bar{F}\left(W_i, h\right) \quad i = 1,2,3 \\ C_4 &= F\left(W_4, h\right) \end{aligned} \tag{4}$$

其中 $\bar{F}$ 为空洞卷积操作，W_i 为卷积核。多尺度空洞卷积层主要是利用不同膨胀率的空洞卷积来获取不同尺度的信息，并加以融合。融合多尺度信息可以综合文本的远近距离关系来为文本分类提供参考。卷积后的结果分别使用最大池化获取文本的最大特征，最大特征对文本分类的作用最大，因此使用最大池化保留最大特征能减少计算参数。对多尺度空洞卷积层的输出分别进行池化，并且将池化后的结果进行拼接。池化层的示意图 4 如下所示。

池化层的计算公式为：

$$\begin{cases} S_i = P\left(C_i\right) \quad i = 1,2,3,4 \\ S = \left[S_1, S_2, S_3, S_4\right]^{\mathrm{T}} \end{cases} \tag{5}$$

其中 P 为最大池化操作，并且将池化后的结果利用全连接层将特征空间映射到分类空间，对文本进行分类，全连接层的计算公式为：

$$y = \operatorname{softmax}\left(WS + b\right) \tag{6}$$

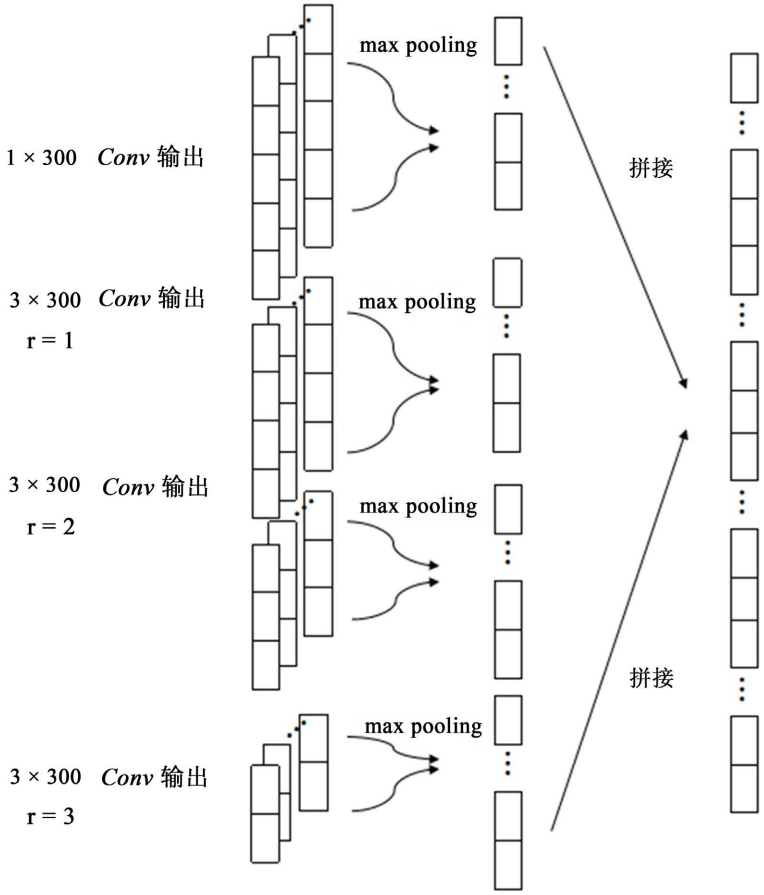

Figure 4. Diagram of pooling layer
图 4. 池化层示意图

其中为W、b全连接计算的参数， y为全连接输出结果。 y是一个维度为类别大小的向量，其中每一列代表该文本被划分到该类别的概率。在预测样本类别的阶段，如果想要预测该文本的类别，则使用公式：

$$\hat{y} = \text{argmax}(y) \tag{7}$$

本文使用的损失函数为使用交叉熵损失函数，计算公式为：

$$Loss = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{M} y_{ij} \ln P_{ij} \tag{8}$$

其中P_{ij}为模型预测第i个样本为第j类的概率， y_{ij}为第i个样本其标签第j类的概率，M代表类别数，其中n为小批量训练样本的数量

4. 实验

4.1. 实验数据集

本文实验所使用的数据集是清华大学开源中文文本分类数据集 THUCNews [18]，该数据集中包含约 84 万篇新闻文档，整个数据集的类别分类和数量如表 1 所示。

从上表可以看出，该数据集的种类数量分布极其不均匀，比如数量最多的类别“体育”与数量最少的类别“星座”其数量相差约 37 倍。样本分布不均衡，可能导致样本少的类别提取到的特征较少，故很

难在其中提取规律，导致分类模型过度依赖样本量多的类别的特征，进而导致过拟合问题。虽然一般可以通过欠采样或过采样来实现样本均衡，但是在训练样本时，仍然会导致过拟合的问题。为保证模型能够均衡的提取样本特征，本文在 14 个类别中选取其中 10 个类别进行实验，其中每个类别选取 5000 训练样本，1000 测试样本，500 验证样本，具体样本类别及数量如表 2 所示。

Table 1. THUCNews category and quantity
表 1. THUCNews 类别和数量

类别	数量	类别	数量
体育	131,604	时尚	13,367
娱乐	92,631	时政	63,085
家居	32,585	游戏	24,372
房产	20,049	科技	162,928
教育	41,935	财经	37,097
股票	154,397	社会	50,848
彩票	7587	星座	3577

Table 2. Category and quantity of data
表 2. 实验数据集类别数量

类别	数量	类别	数量
体育	6500	时尚	6500
娱乐	6500	时政	6500
家居	6500	游戏	6500
房产	6500	科技	6500
教育	6500	财经	6500

本文采用固定长度的预训练向量作为输入，为后续的语义提取层的计算提供方便。文本数据集中的文本长度层次不齐，本文使用 Python 对文本长度进行分析，输出文本长度箱线图如图 5 所示：

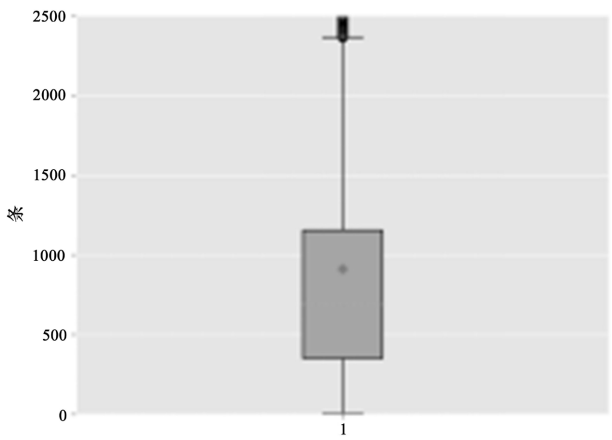

Figure 5. Box chart of text length
图 5. 文本长度箱线图

其中图中的点为均值点。从上图可以看出，有接近一半数据集的文本长度在400~1100左右，因此本文选择将文本长度固定为600。

4.2. 数据预处理

本文选用的为标准文本分类数据集，因而无需删除无效样本进行数据清洗，只需将文本中的英文大小写转换为小写即可。本文采用的停用词词库为哈工大停用词词库，并且调用jieba第三方库的接口对样本进行分词，由于样本中文本的长度层次不齐，为了更好地进行后续操作，本文将文本的长度固定为600。本文选用的预训练词向量为搜狗新闻在大规模语料库预训练好的基于词和bi-gram的Word2vec词向量[19]，词向量的维度为300维，之后建立训练集的词汇表，词汇表的长度为403,277。

4.3. 评价指标

模型训练完成后，一般需要选用模型评价指标来衡量模型的好坏，通常使用精确率(Precision)、召回率(Recall)、F值(F-score)来衡量模型的好坏[18]。以二分类为例，通过预测类别和真实类别的对比形成二分类混淆矩阵，如表3所示。

Table 3. Binary confusion matrix
表3. 二分类混淆矩阵

真实类别	预测类别	
	正例	负例
正例	TP	FN
负例	FP	TN

引入精确率P，召回率R，F_1值作为评价指标，计算公式为：

$$\begin{cases} P = \dfrac{TP}{TP + FP} \\ R = \dfrac{TP}{TP + FN} \\ F_1 = \dfrac{2 \times P \times R}{P + R} \end{cases} \tag{9}$$

上述为二分类问题常用的评价指标，当问题拓展到多分类时，不能单纯考虑某一类别的评价指标的数值，而需要综合考虑所有分类类别的评价指标的数值，从而衡量模型的性能，多分类常用的衡量指标为微平均和宏平均。本文选用宏平均作为评价指标，宏平均计算公式如下所示：

$$\begin{cases} Macro\_P = \dfrac{1}{n}\sum_{i=1}^{n} P_i \\ Macro\_R = \dfrac{1}{n}\sum_{i=1}^{n} R_i \\ Macro\_F = \dfrac{1}{n}\sum_{i=1}^{n} F_i \end{cases} \tag{10}$$

4.4. 模型参数设置

超参数的选择对模型性能有较大的影响，本文选择的超参数如表4所示：

Table 4. Parameters setting
表 4. 参数设置表

参数名称	参数值	参数名称	参数值
本文长度	600	词汇表大小	403,277
词向量维度	300	预训练词向量	sgns.sogou.bigram
LSTM 隐藏神经元	300	空洞卷积膨胀率	[1,2,3]
空洞卷积核尺寸	2 和 3	空洞卷积核数量	128*3
普通卷积核尺寸	1	普通卷积核数量	128
卷积步长	1	padding	valid padding (不填充)
全连接 dropout	0.5	学习率	0.001
优化器	Adam	Epoch	20
激活函数	Leaky ReLu	池化操作	最大池化

4.5. 对比实验

1) textRNN (Recurrent Neural Network for Text Classification with Multi-Task Learning)。该模型使用双向 LSTM 来提取文本的语义信息用于文本分类。

2) textCNN (Convolutional Neural Networks for Sentence Classification)。该模型通过使用不同尺寸的卷积来提取不同长度的短语特征用于文本分类。

3) HAN (Hierarchical attention networks for document classification)。该模型通过分层级对文本进行语义提取来进行文本分类，模型将文本分为句子层级和文档层级，每个层级使用双向 GRU 和 Attention 结构来提取层级语义信息。

4.6. 实验结果分析

4.6.1. 消融实验

为证明多尺度空洞卷积的有效性，本文进行了多次实验。

1) LSTM + DCNN。选用级联空洞卷积代替多尺度空洞卷积。多尺度空洞卷积为分别对 LSTM 的输出进行空洞卷积操作，级联空洞卷积为依次进行空洞卷积操作。

2) LSTM + CNN。使用普通卷积来代替空洞卷积，该模型为分别使用普通卷积对 LSTM 的输出进行操作。

3) textDCNN。只使用空洞卷积而不使用卷积核尺寸为 1 的普通卷积。

4.6.2. 结果分析

模型性能评价指标值分别如表 5 所示：

Table 5. Performance comparison of different models
表 5. 模型性能对比

模型	*Macro_P*	*Macro_R*	*Macro_F*
textCNN	91.61	90.46	89.02
textRNN	96.53	96.33	96.13
HAN	96.83	96.93	97.04

Continued

LSTM + DCNN	96.85	96.82	96.80
LSTM + CNN	97.02	96.98	96.96
textDCNN (卷积核尺寸为 2)	97.30	97.28	97.26
textDCNN (卷积核尺寸为 3)	97.44	97.40	97.39
textPDCNN (卷积核尺寸为 2)	97.47	97.46	97.45
textPDCNN (卷积核尺寸为 3)	97.57	97.54	97.53

从上表可以看出，textPDCNN 的卷积核尺寸为 3 时的模型性能最好，此外通过对比 textDCNN 和 textPDCNN，发现卷积核尺寸为 3 的模型比卷积核尺寸为 2 的模型性能更好，性能提升了 0.12%左右。说明卷积核尺寸较大可以使模型获得较大的感受野，连接文本中较远距离的词，获取文本中较远距离词之间的联系，为文本分类提供更好的依据和特征。

通过对比 textDCNN 和 textPDCNN 的模型性能，发现 textPDCNN 的模型性能比 textDCNN 的模型性能更好，并且性能提升了 0.18%左右。说明增加一个卷积核尺寸为 1 的卷积可以从文本中提取更多的信息，从而获得更好的性能。

通过对比 LSTM + DCNN 与 textDCNN，发现 textDCNN 的性能比 LSTM + DCNN 的性能提升了 0.6%左右，说明多尺度空洞卷积层的特征提取比级联空洞卷积层的特征提取更有效，因为级联空洞卷积的特征提取是不断在原来空洞卷积输出的基础上进行特征提取，而空洞卷积在特征提取的过程中会忽略一些信息，所以级联空洞卷积层原来忽略的信息就无法在后面的层级进行特征提取，从而导致可能会忽略掉一些对分类有帮助的信息。而多尺度空洞卷积层中多个空洞卷积之间互不干扰，各自提取信息，并且之后进行特征融合，可以将多个空洞卷积提取的信息融合在一起，互相补充信息，使得重要信息不会被遗漏。

通过对比 textDCNN 与 LSTM + CNN，发现多尺度空洞卷积层比普通分组卷积层效果更好，说明空洞卷积相较于普通卷积，可以提高模型感受野来获取文本中远距离的依赖关系，从而提高文本分类准确率。

模型的 Macro_P、Macro_R 和 Macro_F 对比图如图 6~8 所示：

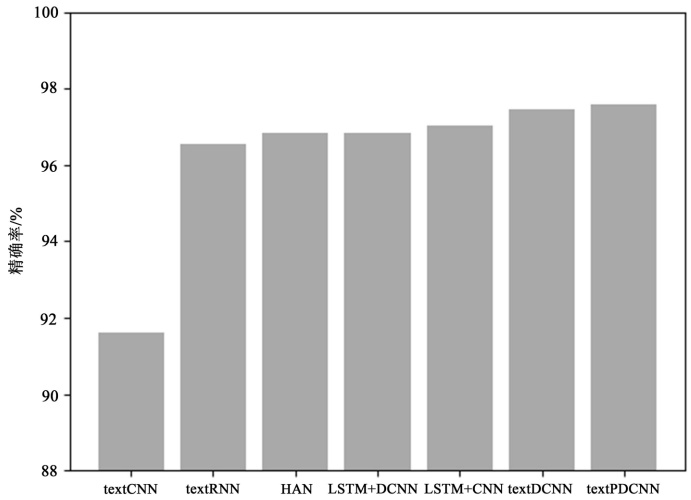

Figure 6. Comparison of *Macro_P*
图 6. 模型 *Macro_P* 对比图

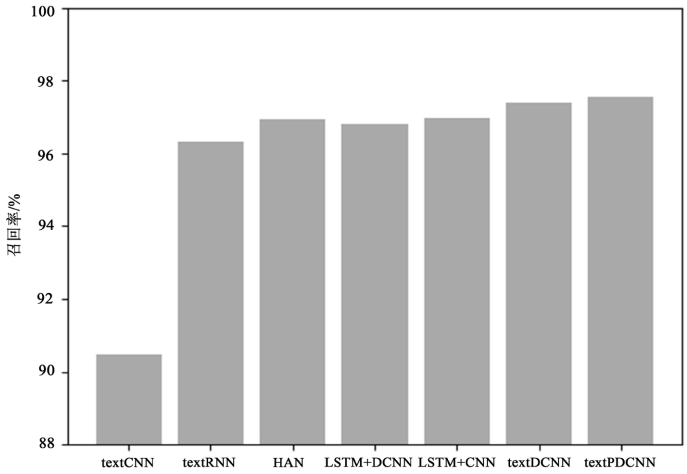

Figure 7. Comparison of *Macro_R*
图 7. 模型 *Macro_R* 对比图

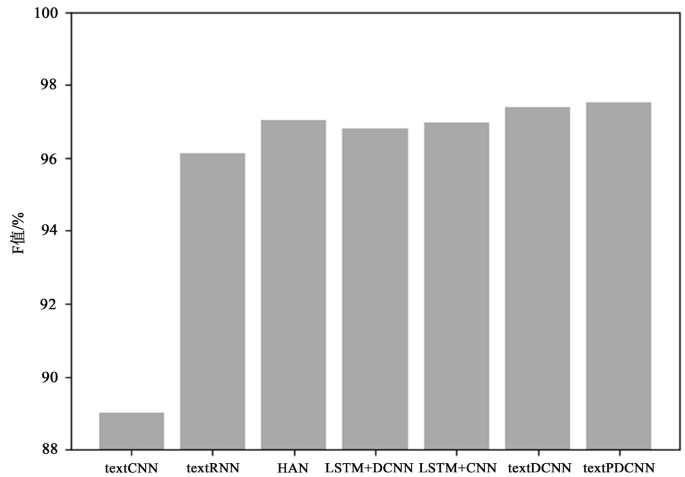

Figure 8. Comparison of *Macro_F*
图 8. 模型 *Macro_F* 对比图

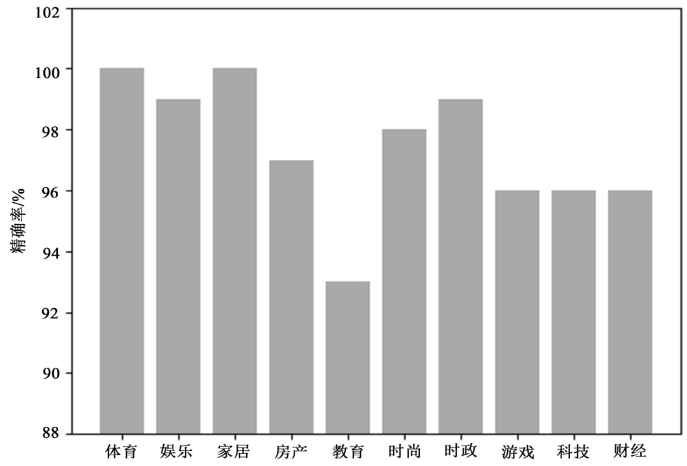

Figure 9. Comparison of the accuracy for different categories
图 9. 不同类别的精确率对比图

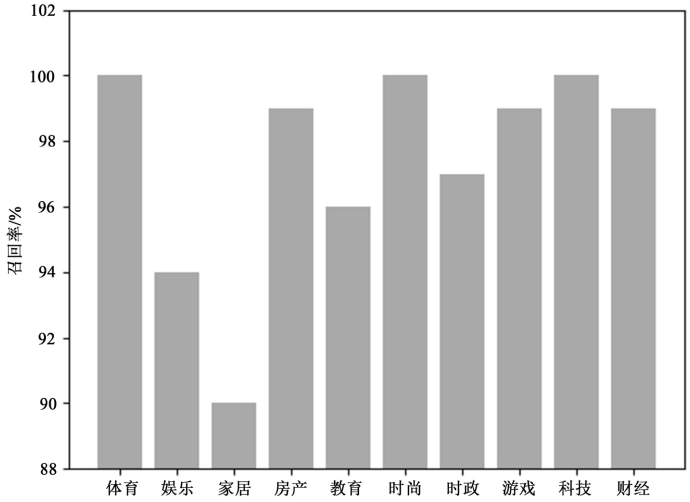

Figure 10. Comparison of recall for different categories
图 10. 不同类别的召回率对比图

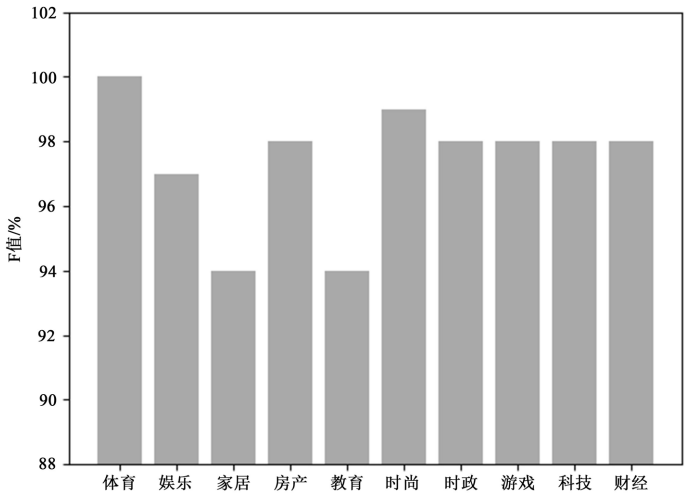

Figure 11. Comparison of F values for different categories
图 11. 不同类别的 F 值对比图

4.6.3. 新闻类别的分类性能对比

模型对不同类别的新闻的分类性能指标如图 9~图 11 所示。

从上图可以看出，“体育”类别的精确率、召回率和 F 值都高达 100%，说明模型能很好的提取“体育”类别的特征进行分类。而“家居”类别和“教育”类别的性能评价指标不是很高，说明模型对这两个类别的特征不能很好的提取。

5. 总结与展望

针对新闻文本分类，本文提出了一种基于空洞空间金字塔池化的模型，该模型结合一维空洞卷积和空洞空间金字塔池化结构来提取文本特征用于分类。模型首先使用 LSTM 来提取文本的完整语义特征，然后使用不同膨胀率的空洞卷积来提取文本不同长度的短语特征，最后使用最大池化来提取文本的显著特征用于分类。实验结果表明，该方法在公共数据集上能够取得不错的效果，为后续新闻文本分类研究提供新的方向和思路。

参考文献

[1] 凤丽洲. 文本分类关键技术及应用研究[D]: [博士学位论文]. 长春: 吉林大学, 2015.

[2] 张小花. 基于文本分类技术的垃圾邮件过滤研究[D]: [硕士学位论文]. 合肥: 安徽大学, 2017.

[3] 何炎祥. 用于微博情感分析的一种情感语义增强的深度学习模型[J]. 计算机学报, 2017, 40(4): 773-790.

[4] 陶文静. 基于卷积神经网络的新闻文本分类研究[D]: [硕士学位论文]. 北京: 北京交通大学, 2019.

[5] 靳小波. 文本分类综述[J]. 自动化博览, 2006, 23(z1): 24-29.

[6] 郭诗瑶. 融合上下文信息的文本分类算法研究及应用[D]: [硕士学位论文]. 北京: 中国邮电大学, 2019.

[7] 刘月. 基于嵌套 LSTM 的中文新闻文本分类研究[D]: [硕士学位论文]. 成都: 西南交通大学, 2019.

[8] 唐雪涛. 基于神经网络嵌入模型的中文文本分类方法研究[D]: [硕士学位论文]. 合肥: 合肥工业大学, 2020.

[9] 王煜. 基于决策树和 K 最近邻算法的文本分类研究[D]: [博士学位论文]. 天津: 天津大学管理学院, 2006.

[10] 赵琳. 基于关联规则的文本类投诉信息分类方法及分类器构建[D]: [硕士学位论文]. 长春: 东北师范大学, 2014.

[11] Joachims, T. (1998) Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Springer, Berlin. https://doi.org/10.1007/BFb0026683

[12] Lewis, D.D. (1998) Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. 10*th European Conference on Machine Learning*, Chemnitz, 21-23 April 1998, 4-15. https://doi.org/10.1007/BFb0026666

[13] Yang, Y.M. (1999) An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval*, **1**, 69-90. https://doi.org/10.1023/A:1009982220290

[14] 姜诚. 基于深度学习的中文长文本分类算法的研究与实现[D]: [硕士学位论文]. 北京: 中国科学院大学, 2020.

[15] Liu, P.F., Qiu, X.P. and Huang, X.G. (2016) Recurrent Neural Network for Text Classification with Multi-Task Learning. *Proceedings of the* 25*th International Joint Conference on Artificial Intelligence*, New York, 9-15 July 2016, 2873-2879.

[16] Kim, Y. (2014) Convolutional Neural Networks for Sentence Classification. *Proceedings of the Empirical Methods in Natural Language Processing*, Doha, 25-29 October 2014, 1746-1751. https://doi.org/10.3115/v1/D14-1181

[17] Yang, Z., Yang, D. and Dyer, C. (2016) Hierarchical Attention Networks for Document Classification. *Proceedings of the* 2016 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, San Diego, June 2016, 1480-1489. https://doi.org/10.18653/v1/N16-1174

[18] Sun, M.S., Li, J.G. and Guo, Z.P. (2016) THUCTC: An Efficient Chinese Text Classifier.

[19] Shen, L., Zhe, E.Z. and Hu, R.F. (2018) Analogical Reasoning on Chinese Morphological and Semantic Relations. Association for Computational Linguistics, Melbourne.