基于语义理解注意力神经网络的多元特征融合中文文本分类

谢金宝 侯永进* 康守强 李佰蔚 张 霄
(哈尔滨理工大学电气与电子工程学院 哈尔滨 150080)

摘 要：在中文文本分类任务中，针对重要特征在中文文本中位置分布分散、稀疏的问题，以及不同文本特征对文本类别识别贡献不同的问题，该文提出一种基于语义理解的注意力神经网络、长短期记忆网络(LSTM)与卷积神经网络(CNN)的多元特征融合中文文本分类模型(3CLA)。模型首先通过文本预处理将中文文本分词、向量化。然后，通过嵌入层分别经过 CNN 通路、LSTM 通路和注意力算法模型通路以提取不同层次、具有不同特点的文本特征。最终，文本特征经融合层融合后，由 softmax 分类器进行分类。基于中文语料进行了文本分类实验。实验结果表明，相较于 CNN 结构模型与 LSTM 结构模型，提出的算法模型对中文文本类别的识别能力最多提升约 8%。

关键词：中文文本分类；多元特征融合；注意力算法；长短期记忆网络；卷积神经网络
中图分类号： TP391.1 **文献标识码**： A **文章编号**：1009-5896(2018)05-1258-08
DOI: 10.11999/JEIT170815

Multi-feature Fusion Based on Semantic Understanding Attention Neural Network for Chinese Text Categorization

XIE Jinbao HOU Yongjin KANG Shouqiang LI Baiwei ZHANG Xiao
(*School of Electrical and Electronic Engineering*, *Harbin University of Science and Technology*, *Harbin* 150080, *China*)

Abstract: In Chinese text categorization tasks, the locations of the important features in the Chinese texts are disperse and sparse, and the different characteristics of Chinese texts contributes differently for the recognition of their categories. In order to solve the above problems, this paper proposes a multi-feature fusion model Three Convolutional neural network paths and Long short term memory path fused with Attention neural network path (3CLA) for Chinese text categorization, which is based on Convolutional Neural Network (CNN), Long Short Term Memory (LSTM) and semantic understanding attention neural networks. The model first uses text preprocessing to finish the segmentation and vectorization of the Chinese text. Then, through the embedding layer, the input data are sent to the CNN path, the LSTM path and the attention path respectively to extract text features of different levels and different characteristics. Finally, the text features are fused by the fusion layer and classified by the classifier. Based on the Chinese corpus, the text classification experiment is carried out. The results of the experiments show that compared with the CNN structure model and the LSTM structure model, the proposed algorithm model improves the recognition ability of Chinese text categories by up to about 8%.

Key words: Chinese text categorization; Multi-feature fusion; Attention algorithm; Long Short Term Memory (LSTM) network; Convolutional Neural Network (CNN)

1 引言

中文文本分类是高效管理与挖掘互联网上海量中文文本信息的重要手段，是自然语言处理中的一个重要的研究方向。中文文本分类的核心任务是通过对中文文本特征的提取与识别来分辨文本类别。深度学习算法，相较于传统的基于统计及人工设计公式的特征提取方式，能自主学习文本的深层语义特征[1]。多层卷积神经网络(Convolutional Neural Network, CNN)适合用来提取文本的深层特征。为了更全面地提取特征，通过采用多个卷积核尺寸不同的 CNN 通路获得文本中对应不同尺寸卷积核的文本特征[2]，然而经过卷积层提取的特征向量经过池化、全链接后就进入分类器无法充分体现特征之间的深层语义联系。与之相比，在采用多尺寸卷积核提取文本特征时，使用尺寸为 3 倍与 5 倍词向量维度的卷积核分别提取文本特征，对提取的特征向量进行融合处理后再进入分类模块实现文本分类，可以获得更好的效果[3]。输入模型的文本向量的利用方

收稿日期：2017-08-17；改回日期：2018-01-15；网络出版：2018-03-15
*通信作者：侯永进 hyj1516@163.com
基金项目：黑龙江省海外学人基金(1253HQ019)
Foundation Item: The Overseas Scholars Fund Project of Heilongjiang Province (1253HQ019)

式也会影响模型分类效果。比如，在对文本进行情感分类时，结合情感词典分别获得原始语料的连续特征和二值特征作为输入，然后分别将二者通过 CNN 提取特征进行后续处理，来获得原始文本不同角度特征[4]。文本处理粒度上，使用 CNN 分析微博情感倾向时，字级词向量效果优于词级词向量[5]。基于 CNN 对英文文本分类时，对构成文本的句子进行标注，以句子为粒度进行文本表示取得较好的效果[6]。两者的主要区别在于语料语言不同及文本表示的粒度不同并且后者[6]采用了具有多尺寸卷积核结构的 CNN。

相较于 CNN，长短期记忆(Long Short Term Memory, LSTM)网络能体现相隔较远的两个文本元素之间的相互联系。LSTM 具有记忆能力且适合处理序列数据，被应用于文本预测[7]、情感分类[8]等任务中并取得很好效果。LSTM 对文本向量的处理的局限性在于以链状结构组成阵列时，对文本的深层特征的挖掘能力不强[9]。可以采用树形结构 LSTM 配合情感极性转移[9]，或采用双向 LSTM[10]来挖掘文本的深层语义信息。

文本中的重点内容分布不均。注意力机制对每个文本元素的语义编码配以不同的权重，来区分其对文本识别重要性的大小。以 LSTM 做编码器，配合注意力机制可以实现一种具有对文本重点内容的注意力与识别能力的机器阅读模拟器[11]。现在有研究者将模仿人类的阅读行为的注意力机制应用于识别句子的情感倾向[12]、多层次并行化处理文本[13]、字符级的问答系统[14]和机器翻译[15]等前沿自然语言处理任务。

为了更加全面地挖掘文本特征，通过采用多元特征融合的方法将 CNN, LSTM 和注意力机制算法结合起来，本文提出一种适合中文文本分类任务的多元特征融合的神经网络模型(3CLA)。

2　神经网络通路

2.1 CNN 通路

经过文本预处理后，中文文本被表示成序列化的数据形式，所以 CNN 通路的卷积层均采用 1 维卷积。通过调整卷积核尺寸可以获得数据在不同宽度视野下的局部特征。为了包含最小粒度字词的上下文，卷积核寸依次选为 3, 4, 5 倍的词向量维度。同一条 CNN 通路中卷积核的尺寸相同，池化层均采用最大池化。通过将不同卷积核的 CNN 的处理结果拼接在一起，以更全面、更细致地挖掘深层特征[3]。3CNN 通路神经网络结构示意图如图 1 所示。

嵌入层作为输入层，每行为一篇文档的向量，依次经过 3 种 3 层 1 维卷积层提取特征和最大池化

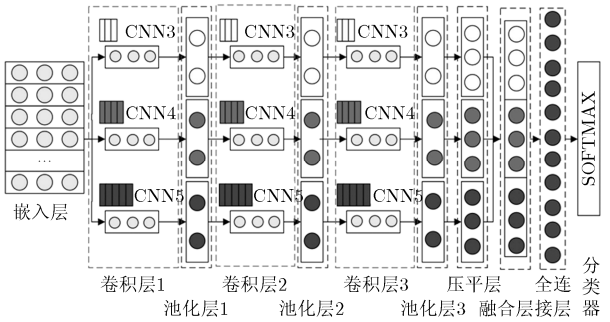

图 1　3CNN 通路神经网络结构示意图

降维，CNN3, CNN4, CNN5, 3 条通路的输出经压平层(flatten)压为 1 维向量，经融合以及全连接后进入分类器。

2.2 LSTM 通路

由 LSTM 单元连接组成的循环神经网络层适合用来处理序列形式的文本数据，并产生序列化的输出数据。由于每一个 LSTM 单元都具有记忆能力，并能控制在合适的时机释放存储的状态数据从而能够体现出文本中相距较远的两个词之间的相互影响。LSTM 的基本外部结构与神经元的连接方式如图 2 所示。

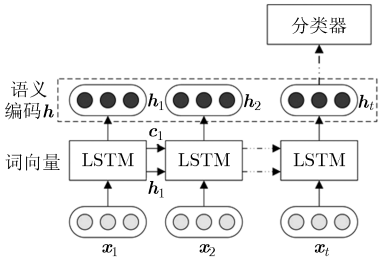

图 2 基于 LSTM 的文本分类流程图

图 2 中 LSTM 阵列中的虚线箭头代表第 3 至第 $t-1$ 个 LSTM 单元，$\boldsymbol{x}_t$ 为当前时间步的输入。其中，$\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_t \in R^{\mathrm{Id}\times 1}$, Id 为输入数据的维度。$\boldsymbol{h}_t$ 为当前时间步的 LSTM 隐层状态，$\boldsymbol{h} \in R^{n\times 1}$, n 为隐层维度。由语义编码到分类器处的虚线箭头代指由 LSTM 输出到分类器之间的数据还会经过激活层、全链接层等中间层次的处理。每一个 LSTM 单元包含遗忘层、输入层、细胞状态层和输出层，赋予了 LSTM 单元判断力、控制力与记忆力。处于 LSTM 单元阵列中的每个单元还会接收与传递细胞状态与隐层状态。

遗忘层更新计算公式如式(1)所示，其中 σ 为 sigmoid 函数。

$$\boldsymbol{F}_t = \sigma\left(\boldsymbol{W}_F \bullet [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{B}_F\right) \tag{1}$$

输入层更新计算公式如式(2)所示：

$$\boldsymbol{I}_t = \sigma\left(\boldsymbol{W}_I \bullet \left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{B}_I\right) \tag{2}$$

LSTM 单元状态更新公式如式(3)所示：

$$\boldsymbol{c}_t = \boldsymbol{F}_t \odot \boldsymbol{c}_{t-1} + \left(1 - \boldsymbol{F}_t\right) \odot \tilde{\boldsymbol{c}}_t \tag{3}$$

$$\tilde{\boldsymbol{c}} = \tanh\left(\boldsymbol{W}_C \bullet \left[\boldsymbol{h}_{t-1}, \boldsymbol{X}_t\right] + \boldsymbol{B}_C\right) \tag{4}$$

单元输出如式(5)所示：

$$\boldsymbol{O}_t = \sigma\left(\boldsymbol{W}_O \left[\boldsymbol{h}_{t-1}, \boldsymbol{X}_t\right] + \boldsymbol{B}_O\right) \tag{5}$$

当前时间步隐层状态更新公式为

$$\boldsymbol{h}_t = \boldsymbol{O}_t \odot \tanh\left(\boldsymbol{c}_t\right) \tag{6}$$

上述公式中，遗忘层权重矩阵 $\boldsymbol{W}_F$、输入层权重矩阵 $\boldsymbol{W}_I$、LSTM 单元状态权重矩阵 $\boldsymbol{W}_C$ 以及输出层权重 $\boldsymbol{W}_O$ 的形式均属于 $\boldsymbol{R}^{\mathrm{Id}\times\mathrm{Od}}$，其中 Id 为输入维度，Od 为输出维度。

2.3 基于语义理解的注意力模型

本文提出一种适用于中文文本篇章分类的“语义理解注意力神经网络模型”。人类在阅读中文文章时，并非照片式记忆，而是由对构成中文文本的基本元素即字、词和标点的基本视觉感知与理解，然后结合语法、使用习惯以及上下文之间的联系对基础元素的学习进行非线性融合，连词成句，连句成文。从文章开头到结尾的顺序阅读的过程中，通过文章中的字、词与其它与其相关的字、词之间语义联系的理解，读者会产生针对文章重点内容的注意力，在接下来阅读后文时会对与文章重点相关性较大的内容予以重点关注，相当于对重点内容配以更大的注意力权重。本文提出一种基于语义理解的注意力算法模型如图 3 所示。

该模型模仿人类在阅读中文文本时注意力的产生过程，首先以 LSTM 作为编码器对经过文本预处理，已经被向量化表示的中文文本进行编码，获得文本的语义编码。然后对当前时间步的语义编码与前一时间步的语义编码进行初步的学习，接下来对初步学习的结果进行融合与学习产生注意力权重分布。最终通过 softmax 将注意力权重分布归一化成概率形式并被分配到 LSTM 的各个隐层状态上，用于分类。

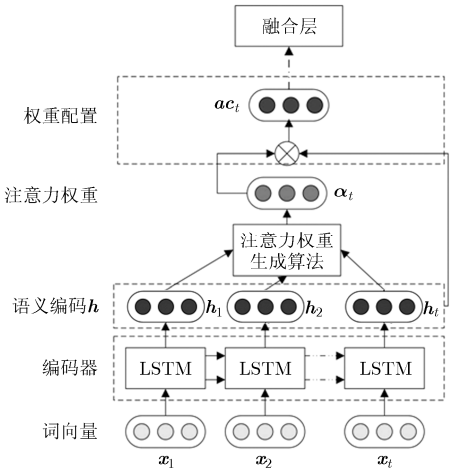

图 3 注意力算法通路流程图

注意力通路模型中的编码器选用 LSTM，这里的 LSTM 阵列仅用于注意力内部做编码器，与 LSTM 通路中的 LSTM 单元不同且互不影响。

注意力权重生成的内在机制如图 4 所示。

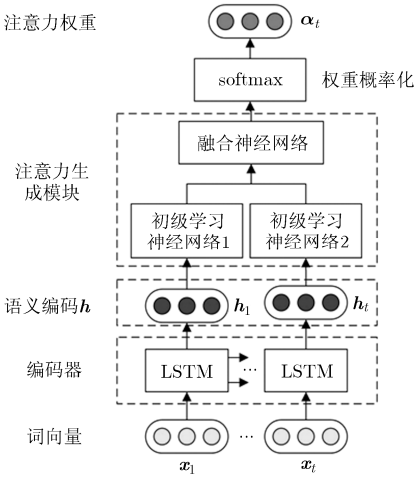

图 4 注意力权重生成原理图

定义算法模型总的输出如式(7)所示：

$$p\left(\boldsymbol{y}_t \left| \boldsymbol{y}_{t-1}, \boldsymbol{X}\right.\right) = g\left(\boldsymbol{h}_t, \boldsymbol{ac}_t\right) \tag{7}$$

其中，$\boldsymbol{X}$ 为当前时间步的输入，其为 1 维数据，如式(8)所示：

$$\boldsymbol{X} = \left[\boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_n\right] \tag{8}$$

作为编码器的 LSTM 阵列，其隐层状态如式(9)所示：

$$\boldsymbol{h}_t = f\left(\boldsymbol{U}_1 \boldsymbol{x}_t + \boldsymbol{W}_1 \boldsymbol{h}_{t-1}\right) \tag{9}$$

式中，$\boldsymbol{U}_1 \in R^{\mathrm{Id}\times\mathrm{Od}}$，$\boldsymbol{W}_1 \in R^{\mathrm{Id}\times\mathrm{Od}}$，Id 为输入维度，Od 为输出维度。

经过 LSTM 阵列对输入进行编码生成的语义编码 $\boldsymbol{h}_t$ 进入内在机制如图 4 所示的注意力权重生成模块。首先由“初级学习神经网络”对当前时间步的隐层状态 $\boldsymbol{h}_t$ 及对前一时间步的隐层状态 $\boldsymbol{h}_{t-1}$ 分别进行初步学习，结果分别如式(10)与式(11)所示。

$$\boldsymbol{m} = \tanh\left(\boldsymbol{h}_t \bullet \boldsymbol{U}_m + \boldsymbol{b}_m\right) \tag{10}$$

$$\boldsymbol{d} = \tanh\left(\boldsymbol{h}_{t-1} \bullet \boldsymbol{W}_d + \boldsymbol{b}_d\right) \tag{11}$$

接下来式(10)与式(11)的输出 $\boldsymbol{m}$ 与 $\boldsymbol{d}$ 进入式(12)对应的图 4 中的融合神经网络中进行融合与学习获得注意力权重 $\boldsymbol{e}_t$，如式(12)所示：

$$\boldsymbol{e}_t = \exp\left(\left(\boldsymbol{m} + \boldsymbol{d}\right) \bullet \boldsymbol{W}_e + \boldsymbol{b}_e\right) \tag{12}$$

对注意力权重 $\boldsymbol{e}_t$ 经过 softmax 归一化处理为概率形式的注意力权重，如式(13)所示：

$$\boldsymbol{a}_t = \frac{\exp(\boldsymbol{e}_t)}{\sum_{j=1}^{T_x}\exp(\boldsymbol{e}_k)} \tag{13}$$

最后进行注意力权重的分配，如式(14)所示：

$$\boldsymbol{ac}_t = \boldsymbol{h} * \boldsymbol{a}_t \tag{14}$$

其中，$\boldsymbol{h} = \left[\boldsymbol{h}_1\ \boldsymbol{h}_2\ \cdots\ \boldsymbol{h}_{T_x}\right]$，最后返回 $\boldsymbol{ac}$ 进入融合层，参与后续处理。以上公式中，$\boldsymbol{U}_t, \boldsymbol{W}_a, \boldsymbol{W}_h \in R^{\mathrm{Od}\times\mathrm{Od}}$，“$\circ$”代表矩阵点积运算。

3　基于多元特征融合的中文文本分类模型

3.1 模型总体介绍

通过多元特征融合的方式可以更加全面地挖掘文本特征，提高系统对文本类别的识别能力。首先，引入“通路”的概念。“通路”是指以一种核心神经网络算法构成的一条数据通路。模型中包含 3 种神经网络组成的通路：CNN 通路、LSTM 通路、注意力算法通路。其中 3CNN 通路由 CNN3, CNN4, CNN5 并联组成。3 种通路基于各自核心神经网络结构的功能与结构特点，能够提取文本数据不同方面的特征，在多元特征融合中发挥各自的作用。3 条 CNN 通路可以提取输入数据在 3 种卷积核粒度下的特征。LSTM 通路提取的输入数据特征可以体现出文本上下文数据之间的联系与影响。注意力算法通路(attention)，产生注意力权重，提高模型对数据中重要内容的识别能力。3 种通路可以根据实际需要进行增减及组合。

基于多元特征融合的中文文本分类模型如图 5 所示。

主要步骤分为 3 个部分：文本预处理、多元特征融合、分类。CNN3, CNN4, CNN5 组成 3CNN 通路，结构均与前文所述的 CNN 通路结构一致。LSTM 通路为单层的 LSTM 网络。

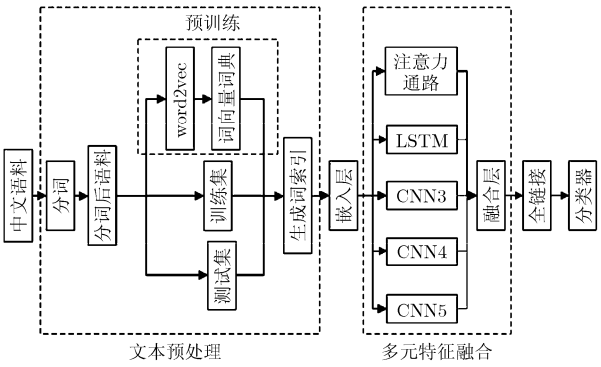

图 5 多元特征融合中文文本分类模型

3.2 文本预处理

原始的中文文本需要通过文本预处理过程来转化成便于计算机处理的数字形式。数字化的过程中要尽可能地保留原始中文文本的特征及语义信息，尽量减少信息损失。文本预处理的主要步骤如下：

(1)分词：中文文本固有特点之一是字与字之间、字与词之间的区分主要依赖语义与语境。与英文不同，中文文本字与字之间没有空格。首先要对中文文本原始语料使用 jieba 软件进行分词。为了充分保留文本信息，分词后的语料依然保留标点与所有字、词。

(2)生成词向量词典：使用 word2vec 算法对分词后的语料先单独进行预训练生成词向量词典。其中的字词是不重复的，word2vec 会对文本中的字、词和标点等基本元素的出现频率进行统计，通过无监督训练，获得作为语料基础构成元素的字词对应的指定维度的向量表征。

(3)数据集划分：在源代码中设定训练集和测试集的划分比例参数为 0.6，将数据集中的 60%的文档用来训练模型，剩余的 40%的文档用来做验证集和测试集，这样划分比较能说明模型性能与泛化能力。

(4)文本数字化：将分词后语料文本中的每个独立的字、词和标点和步骤(2)中生成的词向量词典进行比对，获得其索引，即字、词和标点在向量词典中的序号，这样整篇中文文档都以整数索引序号表示从而实现索引形式的数字化。这样处理有利于降低文本表示的数据维度，节约计算资源，提高处理效率。

(5)文本向量化：在数据进入到神经网络训练前需要按照词的索引序号从(2)中的词向量词典取出其对应的向量，这样整篇文本被转化为向量的形式。为便于神经网络的训练，以“嵌入层”作为神经网络的输入层，每批训练数据会被组织成一个嵌入矩阵。因为每篇文档长度不统一，这里设定每篇文档中最多取出 1000 个元素(字、词或标点)，长度不足 1000 的通过补 0 来补齐长度。

(6)标签：所用原始语料，全部 9833 文档共分 20 类，分别处于 20 个文件夹下。通过区分取出每篇文档的不同路径赋予文档 one-hot 形式的标签。对应 20 个类别，标签共 20 位。

3.3 多元特征融合

通过多元特征融合的方式可以更加全面细致地提取文本的深层语义特征。在经过 CNN 通路、LSTM 通路和注意力算法通路提取不同层次的文本特征后，通过融合层来实现各个通路输出的融合。融合层采用合并拼接各个神经网络通路输出向量的

方式，将各个通路提取的文本特征向量拼接融合在一起。各条通路的输出均为一维的数据。采用这种融合方式的优点是，不必对各通路输出进行数据进行同一维度上尺度的统一，可以避免数据信息损失。多元特征融合示意图如图 6 所示。图 6 中 CNN 通路、LSTM 通路、注意力算法通路输出的一维向量在融合层进行拼接融合生成向量 $\boldsymbol{M}$ 。融合层的输出 $\boldsymbol{M}$ 如式(15)所示：

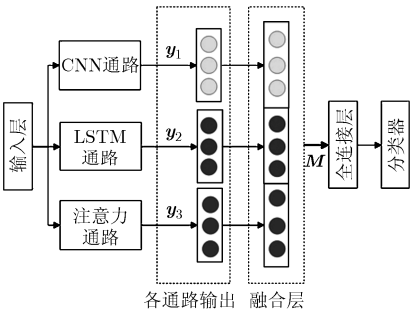

图 6 多元特征融合示意图

$$\boldsymbol{M}=[\boldsymbol{y}_1\ \ \boldsymbol{y}_2\ \ \boldsymbol{y}_3] \tag{15}$$

3.4 分类器

模型所用分类器为 softmax，会对输入进行归一化处理。初始化时需要预先设定数据集类别数，这里所用语料类别数为 20，softmax 输出样本属于这 20 个类别的分数，以其中最大值对应的类别标签作为结果。以测试分类准确率，作为评价通过训练过程获得的模型性能的评价指标。模型的训练与测试主要过程如下：

(1)模型训练：每一条神经网络通路都从嵌入层获得数据。CNN 通路的训练目标是网络参数权重 $\boldsymbol{W}_{cn}$ 和偏置 $\boldsymbol{b}_{cn}$。LSTM 通路的训练目标是权重 $\boldsymbol{W}_{LS}$ 和偏置 $\boldsymbol{b}_{LS}$。注意力通路的训练目标为权重 $\boldsymbol{W}_{AT}$ 和偏置 $\boldsymbol{b}_{AT}$。数据沿神经网络正向传播时，所有通路的输出在融合层进行多元特征的融合。最终的训练目标为：$\boldsymbol{Q}=\{\boldsymbol{W}_{co},\boldsymbol{W}_{LS},\boldsymbol{W}_{AT},\boldsymbol{b}_{co},\boldsymbol{b}_{LS},\boldsymbol{b}_{AT}\}$。反向传播时，以损失函数 Loss 对 $\boldsymbol{Q}$ 求偏导，进行目标参数的更新。反向传播采用的方法是 Adadelta 优化算法。

(2)模型测试：用步骤(1)中训练获得的模型与权重，测试数据沿神经网络正向传播至分类器。分类器输出测试样本属于每一类的概率，其中的最大值对应的序号即为测试样本文档所属类别。

4 实验与结果分析

4.1 实验数据与预处理

中文文本分类实验所采用的实验语料为复旦大学计算机与技术系国际数据库中心自然语言处理小组组织制作。所采用语料集包含中文文档 9833 篇，共分为 20 个种类。实验中以 60%的语料来进行训练，40%来进行测试。实际以 5900 篇做训练集，训练模型待训练参数，剩余的 3933 篇作为验证集与测试集。

首先采用 jieba 软件对原始语料进行分词。分词后的语料按照 6:4 的比例分成训练集与测试集。将分词后的语料作为输入，通过 word2vec 无监督训练获得语料对应的词向量词典。其中字、词的范围与原始语料的字、词和语境的匹配程度对实验结果会产生一定的影响。word2vec 中的参数设置中，所用模型为 cbow 模型，窗口宽度(window)为 8。使用 word2vec 分别训练制作了 50 维、100 维与 200 维词向量的词典用于实验。

4.2 实验环境配置

实验环境：框架：theano, keras；编程语言：python3.4；处理器：GPU。以 theano 做为底层框架，keras 做为顶层框架来实现模型与算法。相比于 CPU，通过配置 cuda 采用 GPU 运行算法，可以提升模型训练速度。

4.3 实验结果与分析

文本预处理过程中，用 word2vec 软件训练获得词向量词典时可以设定词向量的维度。分别对 50, 100, 200 维长度的词向量的条件下以及不同词典规模下的相关模型进行了文本分类实验及对比分析。

3CNN 模型卷积核可视化结果如图 7 所示。采用热图(heatmap)的方式对实现了可视化的工作。全图由上至下共由 4 部分组成：CNN5, CNN4, CNN3 通路第 3 层卷积层卷积核可视化、热图数值-色阶对比图。

需要说明的是，3 幅可视化热图描述的是 3CNN 各卷积通路第 3 层卷积层卷积核训练结果中的第 1 个卷积核训练结果数据的情况。数值大小在−0.1～0.1 之间。每一个带颜色的小格代表卷积核的一个训练结果的数值，颜色越深说明数值越大。反之，颜色越浅数值越小。

CNN5, CNN4, CNN3 通路第 3 层卷积层用来可视化的卷积核尺寸分别为 1×500 维、1×400 维、1×300 维。将其分别以为每行 100 个数值的形式分行展示。

词向量维度为 100 维时的实验结果如表 1 所示。

模型说明：(1)C3 代指 CNN3 通路，为单条 CNN 通路包含 3 层 1 维卷积层，卷积核尺寸均为 3；(2) LSTM 为单层的 LSTM 网络；(3)C3LSTM 为 1 条 CNN3 通路与 1 条单层 LSTM 并联融合组成的模型；

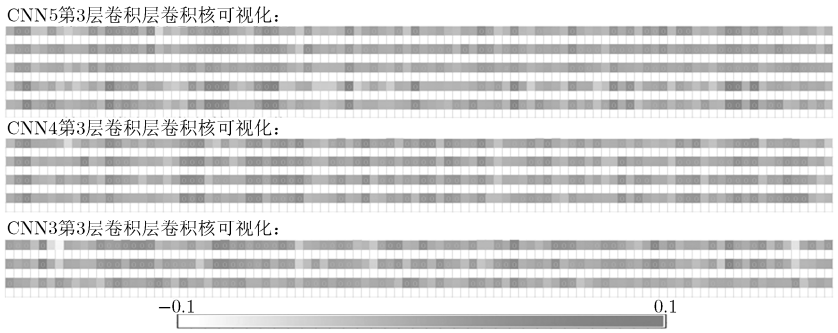

图 7　3CNN 各通路第 3 卷积层卷积核可视化热图

表 1　词向量维度为 100 维时的实验结果

模型名称	C3	LSTM	C3LSTM	LA	C3A	3CNN	C3C5	C3LA	3CL	3CA	3CLA
分类准确率	0.8156	0.8248	0.8289	0.8296	0.8321	0.8576	0.8624	0.8625	0.8753	0.8789	0.8970
平均单轮迭代时间(s)	169.1	579.8	1244.9	4122.2	1563.3	507.0	368.5	2324.7	1529.9	2937.1	3826.1

(4)LA 为 1 条 LSTM 通路与 1 条注意力算法通路并连融合组成的模型；(5)C3A 为 1 条 CNN3 通路与 1 条注意力算法通路并联融合组成的模型；(6) 3CNN 是指 3 条 CNN 通路并联融合的模型；(7)C3C5 为由 CNN3 和 CNN5 并联融合组成的模型；(8)C3LA 为 1 条 CNN3 通路与 1 条 LSTM 通路及注意力算法通路并联融合组成的模型；(9)3CLA 为本文提出的多元特征融合模型。

从实验数据可以看出 LSTM 模型的分类准确率比 C3 提升约 1%，但 LSTM 平均单轮迭代耗时较多。分类准确率 3CNN 相比于 LSTM 提升了约 3%，平均单轮训练耗时与 LSTM 相近。尽管通路数目以及待训练参数的增加，增加了平均单轮迭代训练所需时间，但是由于注意力通路的加入，为整个模型突出了输入数据的重点内容，以及加入 LSTM 后模型可以获得输入数据上下文之间的联系从而提升了模型对文本类别的识别性能，使得分类准确率 3CLA 相比于 3CNN 提升约 4%，而相较于单层的 LSTM 提升约 7%，相较于 C3 提升达到约 8%。

C3A 分类准确率略高于 LA，而 C3A 的平均单论迭代时间远少于 LA。C3A 分类准确率高于 C3LSTM，但是平均单论迭代时间也耗时更多。3CA 与 3CLSTM 相比性能略有提升。与上述各模型相比，本文提出的 3CLA 模型明显提升。

C3C5 是将文献[3]中提出的一种模型思想，映射到本文的实验系统与环境下的一个对比试验。通过实验数据可见，本文提出的 3CLA 相较于 C3C5 测试分类准确率提高了约 3%。原因在于 3CLA 能更全面更细致的挖掘文本特征。

通过以上对实验结果分析，本文提出的多元特征融合模型 3CLA 在分类性能上优于对比模型，证明了本文提出的模型 3CLA 在中文文本分类任务上的有效性，以及 LSTM 和基于语义理解的注意力算法模型为 3CLA 带来了相对于对比模型的中文文本分类性能提升。

对将字词转化为 200 维的向量的情况下多元特征融合模型的性能进行了探索实验。200 维词向量下的目标模型是 C3LA。其在结构上相当于 3CLA 模型去掉卷积核尺寸分别为 4 和 5 的两条卷积通路。

词向量为 200 维时的实验数据如表 2 所示。

表 2　词向量为 200 维时的实验结果

模型名称	C3	LSTM	C3LA
分类准确率	0.8352	0.8469	0.8736
平均单轮迭代时间(s)	364.7	2628.1	6365.8

由实验数据可以看出 LSTM 相较于 C3 模型，分类准确率提升约 1%，单论迭代耗时增加。C3LA 模型相较于 LSTM 分类准确率提升约 3%，相较于 C3 提升约 4%，说明注意力机制与 LSTM 均在发挥作用从而使 C3LA 模型相对于对比模型的中文文本分类性能得到提升，但由于此时单个词向量维度为 200 维，导致模型训练时运算量随通路数量的增加增长较大，从而增加了训练耗时。

与 100 维的词向量表示的情况下相比较，C3, LSTM, C3LA 在 200 维词向量表示下的测试分类准确率均有所提高，因为维度更高的词向量表示可以包含有更丰富的文本语义信息。但时间复杂度增长也很大，因为词向量维度的升高会导致模型中并行

数据维度的急剧上升。

模型 3CLA 在不同条件下的实验结果如表 3 所示。

实验模型说明：3CLA100_100 的含义为 100 维词向量表示下，100%地采用预训练词向量字典时，3CLA 模型的实验。这里的 3CLA100_100 与 100 维词向量下的 3CLA 实验是相同的；3CLA50_100 的含义为 50 维词向量表示下，100%地采用预训练出的词向量字典的条件下，3CLA 模型的实验；3CLA100_80 与 3CLA100_20 的含义为 100 维词向量，采用预训练出的词向量字典的前 80%条件下和采用预训练出的词向量字典的后 20%条件下，3CLA 模型的两个实验。

通过将 3CLA100_100 与 3CLA50_100 的实验数据进行对比可以发现，在同等实验条件下，3CLA 模型在 50 维词向量表示下的，分类性能不及 100 维下的模型表现，因为 50 维词向量所包含的上下文相互影响的语义信息更少，可供提取的文本特征更加有限从而限制了模型对文本类别的识别能力。

通过比较 3CLA100_100, 3CLA100_80 及 3CLA100_20，可以看出 3CLA100_80 相比于 3CLA100_100 分类准确率略有降低，而 3CLA100_20 相比于 3CLA100_100 准确率的下降幅度更大。这是因为文本预处理时获得的词典对训练与测试文本语料、语境的涵盖程度，对同一系统的中文文本分类性能有很大的影响。

通过实验可以看出多元特征融合的模型相对于单一种类神经网络的模型，对中文文本的分类性能更好。通过 100 维词向量下的实验，证明了 3CLA 模型在中文文本分类任务中相对于对比模型的性能优势与有效性。通过综合 100 维与 200 维词向量下的实验结果证明了 LSTM 与基于语义理解注意力算法模型的加入使得模型的中文文本分类能力得到提高，证明了算法的有效性。通过 100 维与 50 维词向量下 3CLA 的性能对比，探索了词向量维度对模型的影响。通过不同词典规模下的对比，探索了词向量词典大小对模型性能的影响。

5 结束语

本文提出一种针对中文文本分类的基于语义理解注意力神经网络同时结合 CNN 与 LSTM 的多元特征融合中文文本分类模型(3CLA)。通过 3 条 CNN 通路充分挖掘文本数据在 3 种不同尺寸卷积核粒度下的特征；同时，通过融合 LSTM 来体现文本数据之间的相互联系；通过融合所提出的基于语义理解的注意力算法模型使相对重要的数据特征在中文文本类别识别过程中发挥更大的作用。将三者融合来实现文本不同层次特征的多元特征融合，以充分而全面地挖掘文本特征。从而提高模型 3CLA 整体对中文文本类别的识别能力。实验结果表明，同等实验条件下，相比于 CNN 结构模型与 LSTM 结构模型，3CLA 模型的中文文本分类性能明显提高。在后续研究中，还会对模型的算法、结构与参数设置进行优化和改进，以进一步提高模型对中文文本识别能力以及减少训练时间。

表 3 模型 3CLA 在不同条件下的实验结果

模型名称	3CLA100_100	3CLA50_100	3CLA100_80	3CLA100_20
分类准确率	0.8970	0.8672	0.89143	0.72133
平均单轮迭代时间(s)	3826.1	1694.6	2822.1	3051.6

参 考 文 献

[1] 孙晓, 彭晓琪, 胡敏, 等. 基于多维扩展特征与深度学习的微博短文本情感分析[J]. 电子与信息学报, 2017, 39(9): 2048–2055. doi: 10.11999/JEIT160975.
SUN Xiao, PENG Xiaoqi, HU Min, *et al*. Extended multi-modality features and deep learning based Microblog short text sentiment analysis[J]. *Journal of Electronics & Information Technology*, 2017, 39(9): 2048–2055. doi: 10.11999/JEIT160975.

[2] KIM Yoon. Convolutional neural networks for sentence classification[C]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014: 1746–1751.

[3] PHAM Ngocquan, KRUSZEWSKI German, and BOLEDA Gemma. Convolutional neural network language models[C]. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin Texas, 2016: 1153–1162.

[4] 陈钊, 徐睿峰, 桂林, 等. 结合卷积神经网络和词语情感序列特征的中文情感分析[J]. 中文信息学报, 2015, 29(6): 172–178. doi: 10.3969/j.issn.1003-0077.2015.06.023.
CHEN Zhao, XU Ruifeng, GUI Lin, *et al*. Combining convolutional neural networks and word sentiment sequence features for Chinese text sentiment analysis[J]. *Journal of Chinese Information Processing*, 2015, 29(6): 172–178. doi:

10.3969/j.issn.1003-0077.2015.06.023.

[5] 刘龙飞, 杨亮, 张绍武, 等. 基于卷积神经网络的微博情感倾向性分析[J]. 中文信息学报, 2015, 29(6): 159–165. doi: 10.3969/j.issn.1003-0077.2015.06.021.
LIU Longfei, YANG Liang, ZHANG Shaowu, *et al*. Convolutional neural networks for Chinese micro-blog sentiment analysis[J]. *Journal of Chinese Information Processing*, 2015, 29(6): 159–165. doi: 10.3969/j.issn.1003-0077.2015.06.021.

[6] ZHANG Ye, MARSHALL Iain, and WALLACE B C. Rational-augmented convolutional neural networks for text classification[C]. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin Texas, 2016: 759–804.

[7] MIYAMOTO Yasumasa and CHO Kyunghyun. Gated word-character recurrent language model[C]. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin Texas, 2016: 1992–1997.

[8] TANG Duyu, QIN Bing, LIU Ting, *et al*. Document modeling with gated recurrent neural network for sentiment classification[C]. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Austin Texas, 2015: 1422–1432.

[9] 梁军, 柴玉梅, 原慧斌, 等. 基于极性转移和 LSTM 递归网络的情感分析[J]. 中文信息学报, 2015, 29(5): 160–167. doi: 10.3969/j.issn.1003-0077.2015.05.020.
LIANG Jun, CHAI Yumei, YUAN Huibin, *et al*. Polarity shifting and LSTM based recursive networks for sentiment analysis[J]. *Journal of Chinese Information Processing*, 2015, 29(5): 160–167. doi: 10.3969/j.issn.1003-0077.2015.05.020.

[10] 刘飞龙, 郝文宁, 陈刚, 等. 基于双线性函数注意力 Bi-LSTM 模型的机器阅读理解[J]. 计算机科学, 2017, 44(6A): 92–96. doi: 10.11896/j.issn.1002-137X.2017.6A.019.
LIU Feilong, HAO Wenning, CHEN Gang, *et al*. Attention of bilinear function based Bi-LSTM model for machine reading comprehension[J]. *Computer Science*, 2017, 44(6A): 92–96. doi: 10.11896/j.issn.1002-137X.2017.6A.019.

[11] CHENG Jianpeng, DONG Li, and LAPATA Mirella. Long short-term memory-networks for machine reading[C]. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin Texas, 2016: 551–561.

[12] WANG Yequan, HUANG Minlie, ZHAO Li, *et al*. Attention-based LSTM for aspect-level sentiment classification[C]. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin Texas, 2016: 606–615.

[13] PARIKH Ankur P, TAKSTROM Oscar, DAS Dipanjan, *et al*. A decomposable attention model for natural language inference[C]. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin Texas, 2016: 2249–2255.

[14] GOLUB David and HE Xiaodong. Character-level question answering with attention[C]. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin Texas, 2016: 1598–1607.

[15] MI Haitao, WANG Zhiguo, and ITTYCHERIAH Abe. Supervised attentions for neural machine translation[C]. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin Texas, 2016: 2283–2288.

谢金宝：　男，1980 年生，副教授，研究方向为自然语言处理、人工智能.

侯永进：　男，1988 年生，硕士生，研究方向为自然语言处理、人工智能.

康守强：　男，1980 年生，教授，　研究方向为智能诊断、人工智能.

李佰蔚：　女，1989 年生，硕士生，研究方向为自然语言处理、人工智能.

张　霄：　男，1993 年生，硕士生，研究方向为人工智能.