

[文章编号] 1003—4684(2019)04-0063-05

基于深度学习的法院信息文本分类

杨 帆, 陈建峡, 郑吟秋, 黄煜俊, 李 超

(湖北工业大学计算机学院, 湖北 武汉 430068)

[摘 要] 为解决在法院数据信息化过程中,海量的法院文书存在缺乏自动管理分类的问题,提出一种基于字符级卷积神经网络的文本分类模型。模型通过卷积神经网络进行特征提取,能够精确有效地解决文本分类问题。实验结果证明,该模型可以实现在测试集上准确率 99.67% 的分类,且训练用时只有常用循环神经网络算法的 50%。

[关键词] 法院信息文本; 卷积神经网络; 字符级; 深度学习; 文本分类

[中图分类号] TP183 [文献标识码] A

随着法院档案数量飞速增长,使用信息化技术对法院档案进行归档整理,提取有效信息,给司法人员对材料的管理和应用提供了极大的便利。对档案进行准确有效的分类则是最初和关键的一步。

文本分类是自然语言处理领域中一个经典问题,指计算机通过算法判别文本所属的类别。文本分类最早的研究起始于上世纪中期^[1]。在较长的一段时间内,机器学习方法是文本分类研究的主流方法。此时比较经典的分类方法包括:朴素贝叶斯、支持向量机(Support Vector Machine, SVM)等^[2]。这类方法需要人为设计特征,而特征对分类效果有很大的影响,如果人工选择的特征不理想,分类精度会降低。近些年来,深度学习凭借在海量数据下表现出的强大特征提取能力,也成为自然语言处理领域的重要研究方法。

为解决传统文本分类方法抽取特征困难等问题,本文提出一种基于字符级卷积神经网络的文本分类模型 OCCNN(One-hot Character Convolutional Neural Network, OCCNN),用于对法院案件判决书进行分类。该模型方法简单、训练时间短,比其他传统方法更加精确,可有效提升分类效率。

1 相关技术概述

1.1 文本表示

使用神经网络进行特征提取,首先要将文本转换为神经网络可以计算的形式,这一步就是文本表

示。传统的文本表示方法包含有热词模型(One-hot)、词袋模型,而在深度学习中常用的文本表示方法包括 Word2vec 模型, Glove 模型等^[3-5]。

1.1.1 One-hot 模型 最初被广泛使用的文本表示方法是 One-hot 方法。One-hot 方法统计文章中出现的所有单词,并编号。将句子生成对应单词编号处为 1,其它位置为 0 的矩阵^[6]。这种表示方法直观简便,但也有重大的缺陷。首先,生成矩阵的维度就是字典的长度,但有时字典会达到数万级,这样将每个词都用万维的向量来表示,简直是内存的灾难。并且 One-hot 矩阵只能体现单词是否存在于句子中,而忽略了单词之间的顺序和上下文关系。

1.1.2 Distributed representation 模型 现在流行的 Word2vec 模型, Glove 模型属于分布式表示(Distributed representation)方法,不同于 One-hot 这种局部表示方法,分布式表示方法通过神经网络对文本的目标词和上下文建模,使词向量包含丰富的语义信息。但训练模型会花费大量时间,会对实际使用造成不便。

为缩减时间和简化步骤,再加上判决书内容单调,训练集小,不像新闻类数据集会产生巨大的词典。本文选择改良的 One-hot 模型来进行文本表示。

1.2 深度神经网络

随着深度学习的兴起,深度学习在不同学科各个领域都有优秀的研发和应用^[7]。如陈翠平使用

[收稿日期] 2019-03-21

[基金项目] 湖北省科技厅自然科学基金青年面上项目(2017CFB326)

[第一作者] 杨 帆(1994-),女,河南三门峡人,湖北工业大学硕士研究生,研究方向为机器学习,自然语言处理

[通信作者] 陈建峡(1971-),女,湖北武汉人,工学硕士,湖北工业大学副教授,研究方向为机器学习,自然语言处理

深度信念网络进行文本分类^[8]。张春云等通过卷积神经网络来自动调整特征的权重^[9]。赵志宏等对经典卷积神经网络模型 LeNet-5 的结构进行了改进,进行有效的字符识别^[10]。在自然语言处理方面,深度学习也在阅读理解、机器写作、对话系统、机器翻译等各个领域取得极大进展^[11]。

1.2.1 循环神经网络 文本数据中单词并非相互独立,都有语义关联。因为循环神经网络基于序列的结构,可以更好的获取前后文本的信息,从而做出更好的预测。因此,循环神经网络以及长短期记忆网络 LSTM(Long Short-Term Memory)一直是自然语言处理任务中的首选^[12]。例如,王红等基于 LSTM 的语义关系抽取模型,可以提取更多的文本信息,更有效地对句子建模^[13]。

1.2.2 卷积神经网络 卷积神经网络在卷积层通过特征拟合能力来减少权值的数量,有效减轻计算负担。在池化层,通过特征压缩,既可以有效地保留重点的结构信息,同时又能减少数据处理量。因而,卷积神经网络起初常应用于图像识别等领域。

2014 年 Kim 的论文比较系统的将非静态卷积神经网络应用于自然语言处理方面,并在情感分类等任务中有比较好的表现^[14-15]。此后,还有刘敬学使用高速神经网络对卷积神经网络进行改进,结合 LSTM 来处理短文本分类问题,Jin Wang 使用知识驱动的卷积神经网络应用于文本分类等,将 CNN 模型改良并应用于自然语言处理问题的研究^[16-17]。图 1 为一个典型的卷积神经网络模型。

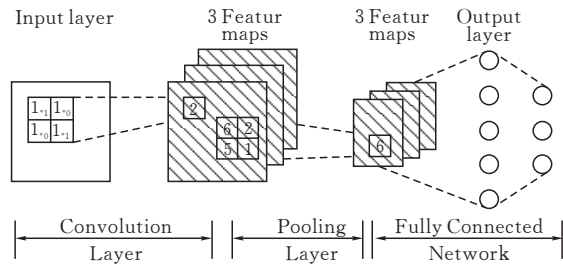


图 1 卷积神经网络(CNN)模型示意图

卷积神经网络要求输入层为固定大小矩阵。图片本身就是以二维矩阵的形式存在,而文本数据则需要在使用文本表示方法进行表示后,再通过卷积神经网络处理。输入部分有单词级及字符级两种方式,单词级别的文本表示方法可以获得更多文本相关信息,可以有效提高应用时的准确性。从字符层面进行文本嵌入,可以提取出高层抽象的文本表示,并且尽可能保存原文信息。

2 基于 OCCNN 的文本分类模型

综上所述,笔者研究并开发了 OCCNN(One-

hot Character Convolutional Neural Network)文本分类模型。首先,将搜集的大量法院判决书文件进行预处理。接着,由训练数据生成词汇表,使用 One-hot 方法进行文本表示,保留字符级信息。然后,使用卷积神经网络自动抽取特征,便于更好地分类预测。

OCCNN 主模型要分为文本预处理、文本表示、神经网络模型训练、文本分类四个部分(图 2)。

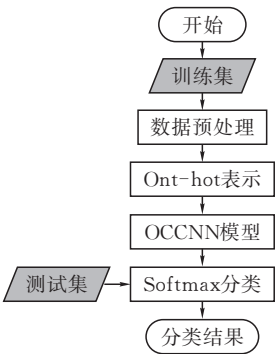


图 2 基于 OCCNN 的文本分类模型

2.1 文本表示

判决书分别为行政、民事、赔偿、刑事和执行判决书,将其修改成统一的格式后,进行简单的预处理,去除多余的符号和无法识别的字符。然后将数据分为训练集、测试集、验证集几个部分。为简化步骤缩短训练时间,本文选择了 One-hot 表示法。

具体步骤为:1)为获得字符级信息,本文选择文本中出现频率最高的 5000 字生成词汇表。2)每一个文本含有词汇表中的字的部分,对应词汇表生成 ID 表。3)设定截取序列长度为 1000,选取 ID 表的 1000 个 ID 生成 One-hot 向量。对应标签也生成 One-hot 向量。将文本矩阵与标签矩阵一一对应,生成输入矩阵。这种方法非常的简单易行,可以省去训练词向量花费的大量时间。图 3 为一个句子按照词汇表拟生成 One-hot 矩阵的过程。

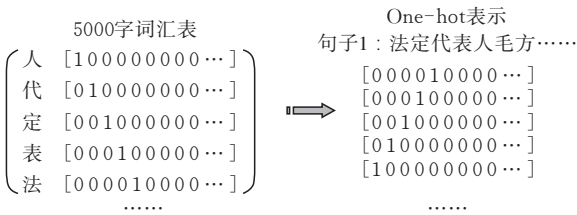


图 3 训练集的 One-hot 表示示意图

2.2 文本分类

本文将文本视为一种字符级别的原始信号,输入卷积神经网络进行处理。使用字符级卷积神经网络构建分类模型并对大量文本进行分类。OCCNN 模型文本分类过程如图 4 所示。

2.2.1 卷积神经网络 在文本表示结束后,进入卷

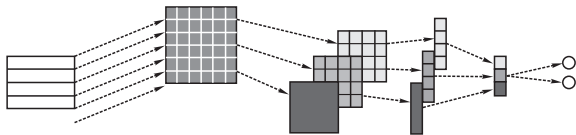


图 4 OCCNN 卷积神经网络模型

积层。如式(1)所示。首先将 One-hot 表示的训练集作为输入矩阵 W_{k+1N} 。然后设定好几个卷积核 $C^{1,a}$ 的参数,其中 1 代表卷积核高度, a 代表向量维度。然后将卷积核放入输入矩阵按照设定好的步长滑动,每滑动到一个位置,将对应元素相乘并求和,最后得到的结果即为卷积得到的特征图。这一步得到和卷积核数量一致的数个特征图 (Feature maps), 将其记做 $h^{1,a}$, $h^{1,a}$ 是输入矩阵与卷积核进行卷积计算再加上偏置 $b^{1,a}$ 的运算和。

$$h^{1,a} = f(x \cdot C^{1,a} + b^{1,a}) \tag{1}$$

卷积运算后,需要通过池化层,也叫做下采样层,来减少权重参数。池化操作在保留重要特征的同时,降低了下一步全连接层的运算量。并且可以有有效的减小过拟合,提高模型的容错性。本文使用的方法是 max-pooling(最大池化),池化计算后输出最大池化操作后的值。然后将池化得到的新特征向量组合,并标明其类别,得到形如 $\{(x_{(m)}, y_{(m)})\}$ 的特征对,其中 $x_{(m)}$ 代表池化后得到的新特征向量, $y_{(m)}$ 则代表文本类别。

池化层后连接全连接神经网络,全连接层后进行 Dropout 操作,然后连接 ReLU 激活。。

2.2.2 Softmax 函数 本文选择 Softmax 函数进行分类。将全连接网络的结果输入 Softmax 函数,将其映射成在区间 $(0,1)$ 之间的值,这些值加和得到 1,将它作为可能为这个类的概率。最后选择概率最大的类,作为最终的分类结果。Softmax 函数的计算公式如下:

$$f(x)_{\varphi} = \frac{1}{1 + \exp(-\varphi^T x)} \tag{2}$$

其中 \exp 代表以 e 为底数的指数函数。

选择梯度下降的方法更新梯度,使用式(3)作为损失函数,损失函数的取值由最小代价函数 $J(\varphi)$ 计算:

$$J(\varphi) = \sum_{i=1}^M y(i) \ln f_{\varphi}(x^{(i)}) \tag{3}$$

softmax 函数返回每一个输出类别的概率,最后通过概率的高低划分文本所属分类类别。

3 实验结果及分析

为了验证 OCCNN 文本分类模型的有效性,本文将数据集在多个文本分类模型上进行分类作为对比实验,对比的基线方法包括两种流行的循环神经

网络模型。

3.1 实验环境设置

实验环境为 Ubuntu16.04 操作系统,使用平台为 Anaconda4. 1. 1 (Python2. 7) 和 TensorFlow (1.4.0)。

3.2 数据源

实验数据采用行政、民事、赔偿、刑事、执行共五个分类的法院判决书作为数据源,数据来自山东省某法院的真实数据。每个分类有 2500 个文本,一共 12500 个文本。每个分类分出 2000 作为训练集, 300 测试集,200 验证集。分类后,文本训练集、测试集、验证集分开进行整合和贴上标签,并进行预处理,去除多余的符号和不可识别的部分后。统一进行格式转换。数据集统计和整合后的各项参数见表 1。

表 1 数据集参数统计表

数据集参数	参数值
文本类别数	5
总数据集大小	12500
文本最大长度	59955
文本平均长度	2687
词典大小	5000
序列长度	1000

本文选择序列长度为 1000,训练集文本的平均长度约为 2000 左右,但由于原始文本中很多字不在词典中,以及文本末尾有很多时间、地点、审判人员以及判决书的固定说辞等对分类无效的信息,将序列长度设置为 2000 不但增加了训练时间,而且并不能增高分类准确率,在多次试验后,本文选择 1000 作为最佳的序列长度。

3.3 实验设置

本文训练时使用字符级卷积神经网络,池化方法选择最大池化,卷积核尺寸为 5,卷积核数目为 128,激活函数选择 ReLU(Rectified linear unit)激活。式(4)为 ReLU 表达式:

$$\text{ReLU} = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases} \tag{4}$$

因为训练集不够大的缘故,训练过程中还是容易发生过拟合。因此在全连接层后设置 dropout 来有效地减少过拟合,设置模型的 dropout 保留比例为 0.5。

本文选择将学习率设为 0.01,学习率太高会导致无法收敛,学习率太低则会极大延长训练时间。将 0.01 设为学习率,实验在三轮迭代后,已经基本收敛。训练至五轮时,准确率已经不再有明显的提升。学习率设置非常合适。实验使用的其他神经网

络参数设置见表 2。

表 2 卷积神经网络参数设置表

参数名	参数含义	参数值
num_classes	类别数	5
num_filters	卷积核数目	256
kernel_size	卷积核尺寸	5
hidden_dim	全链接层神经元	128
learning_rate	学习率	0.01

实验对比所用的基线方法为两个 RNN 改良模型：LSTM 模型和 GRU (Gated Recurrent Unit, GRU)模型。LSTM 克服了 RNN 无法很好处理远距离依赖的问题,而 GRU 则是 LSTM 众多变体中的一个。为了对照实验,基线模型的各项参数尽量与 CNN 模型类似。

表 3 模型分类准确率表

模型分类	CNN			LSTM			GRU		
	precision	recall	f1	precision	recall	f1	precision	recall	f1
行政	1.00	1.00	1.00	0.99	0.97	0.98	0.99	0.98	0.98
民事	1.00	0.98	0.99	0.96	0.94	0.95	0.99	0.95	0.97
赔偿	0.99	1.00	0.99	0.95	1.00	0.97	0.97	0.99	0.98
刑事	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99
执行	1.00	1.00	1.00	1.00	0.99	0.99	0.97	0.99	0.98

从图 5 可以看到,各个神经网络的平均准确率分别为 99.67%,97.8%和 98%。文本使用的 OCCNN 模型相对 LSTM 模型准确率提高 1.87 个百分点,对比 GRU 模型提高了 1.67%,效果最好。其他两个模型分类准确率也很高,但是训练用时则不如 OCCNN 模型优秀。GRU 模型训练时间大约比 OCCNN 模型多出一倍,LSTM 模型则耗时更长。

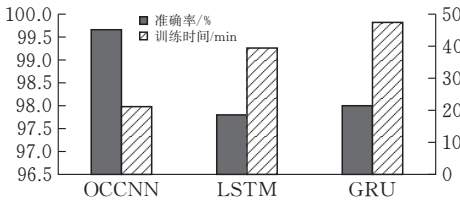


图 5 不同模型分类准确率和训练时间比较

4 结束语

本文提出了一种基于卷积神经网络的 OCCNN 法院判决书分类方法。收集大量法院判决书文本信息,使用 One-hot 方法进行文本表示,快速高效地实现了文本到向量的映射,省去了训练词向量的时间。然后将生成的矩阵引入到卷积神经网络模型训练中进行训练。卷积神经网络的局部连接和权值共享技术,可以有效地提取特征信息,同时还减少了网络参数,大量减少了计算量,并缩短了了训练所需的时间。

3.4 结果分析

实验结果选择准确率(Precision)、召回率(Recall)以及综合评价指标(F_1 -score)作为标准。本文先定义四种分类情况, T_P 该分类判定正确, F_P 非该类被判定为该分类, F_N 该分类被判定为其他类, T_N 非类判定为非该分类。精确率计算式为 $P = \frac{T_P}{T_P + F_P}$,召回率的计算式为 $R = \frac{T_P}{T_P + F_N}$ 。 F_1 值计算式为 $\frac{2}{F_1} = \frac{1}{P} + \frac{1}{R}$ 。其中 F_1 值是召回率和准确率的加权调和平均。

实验得到 OCCNN 模型、和两个比照模型针对各个分类的评价标准计算结果如表 3 所示。

由于实验文本本身的特点,不同分类的文本之间的差别很大,相比新闻等分类素材更容易得到很好的分类效果。对比试验结果证明:本文使用的卷积神经网络模型结构简单且训练速度快,相比传统分类方法,省去了人工选择特征的步骤。和两种循环神经网络模型相比,分类准确率更高,分类使用时间更短,可以高效准确的解决法院判决书的分类问题。

[参 考 文 献]

[1] 吴军.数学之美[M].北京:人民邮电出版社,2014: 10-25.

[2] 杨杰明. 文本分类中文本表示模型和特征选择算法研究[D].长春:吉林大学,2013.

[3] 吴慕雅,魏苗.从深度学习回顾自然语言处理词嵌入方法[J].电脑知识与技术,2016,12(36):184-185.

[4] 周练.Word2vec 的工作原理及应用探究[J].科技情报开发与经济,2015,25(2):145-148.

[5] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[J]. neural information processing systems, 2013: 3111-3119.

[6] 闫琰.基于深度学习的文本表示与分类方法研究[D].北京:北京科技大学,2016.

[7] 尹宝才,王文通,王立春.深度学习研究综述[J].北京工业大学学报,2015,41(1):48-59.

[8] 陈翠平. 基于深度信念网络的文本分类算法[J]. 计算机系统应用, 2015, 24(2): 121-126.

[9] 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.

[10] 赵志宏, 杨绍普, 马增强. 基于卷积神经网络 LeNet-5 的车牌字符识别研究[J]. 系统仿真学报, 2010, 22(3):638-641.

[11] 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.

[12] Greff K, Srivastava R K, Koutnik J, et al. LSTM: A search space odyssey[J]. IEEE Transactions on Neural Networks & Learning Systems, 2016, 28(10):2222-2232.

[13] 王红, 史金钊, 张志伟. 基于注意力机制的 LSTM 的语义关系抽取[J]. 计算机应用研究, 2018(5):1417-1420.

[14] Kim Y. Convolutional neural networks for sentence classification[J]. empirical methods in natural language processing, 2014: 1746-1751.

[15] Kim Y, Jernite Y, Sontag D, et al. Character-aware neural language models[J]. national conference on artificial intelligence, 2016: 2741-2749.

[16] 刘敬学,孟凡荣,周勇,等.字符级卷积神经网络短文本分类算法[J].计算机工程与应用,2019,55(5):135-142.

[17] Wang J, Wang Z, Zhang D, et al. Combining knowledge with deep convolutional neural networks for short text classification[C]. international joint conference on artificial intelligence. Melbourne, Australia, 2017: 2915-2921.

Research on Classification of Court Information
Texts Based on Deep Learning

YANG Fan, CHEN Jianxia, ZHENG Yingqiu, HUANG Yujun, LI Chao
(School of Computer Science, Hubei Univ. of Tech., Wuhan 430068, China)

Abstract: In the process of court data informatization, there is a lack of automatic management classification in massive court documents. This paper proposes a text classification model based on character-level convolutional neural network, which can effectively solve the problem. The model extracts features through convolutional neural networks, which can classify texts efficiently and accurately. Experiments show that the model can achieve an accuracy rate 99.67% of classification on the test set, and the training time is only 50% of the commonly used Recurrent Neural Networks.

Keywords: court information text; convolutional neural network; text classification; character level; deep learning

[责任编辑: 张岩芳]