·专辑:人工智能与情报学·

基于卷积神经网络的文献自动分类研究

郭利敏

(1.上海图书馆 上海 200031)

摘 要:人工智能技术的蓬勃发展,驱动着文献自动分类由基于规则的分类向基于机器学习的方向发展。文章在对深度学习概述的基础上,将卷积神经网络引入到了文献自动分类,构建了基于题名、关键词的多层次卷积神经网络模型,使之能够根据文献的题名和关键词自动给出中图分类号。通过在TensorFlow平台上的深度学习模型,利用《全国报刊索引》约170万条记录进行模型训练,并对7000多篇待加工的文献做中图法分类预测,其在生产情况下一级分类准确率为75.39%,四级准确率为57.61%。当置信度为0.9时,一级正确率为43.98%,错误率为1.96%,四级正确率为25.66%,四级错误率为5.11%。证明该模型有着较低的错误率,可为《全国报刊索引》分类流程的半自动化提供帮助,解决存在的编目人员紧缺、加工质量和效率下降等问题。

关键词:人工智能;智能图书馆;深度学习;卷积神经网络;TensorFlow;自动分类

中图分类号 TP18;G254.11 **文献标识码**:A DOI:10.11968/tsyqb.1003-6938.2017119

Study of Automatic Classification of Literature Based on Convolution Neural Network

Abstract With the rapid development of artificial intelligence, the automatic classification of literature is changing from the rule-based to the machine learning. After an outline of deep learning, the paper introduced convolution neural network into the automatic classification, constructing a multi-level model based on the title and the key words and thus CLC is given automatically. Through the deep learning model in TensorFlow, about 1700000 records of National Newspaper Index were used to make model train. More than 7000literature were processed with the model and the result is: under the production condition, the accuracy of the first classification is 75.39%; the accuracy of the fourth classification is 57.61. When the confidence is 0.9, the correct rate of the first classification is 43.98%, error rate is 1.96%; correct rate of the fourth classification is 25.66%, the error rate is 5.11%.This shows that the model can be used to help realize the semi-automatic in the classification of National Newspaper Index and other problems.

Key words artificial intelligence; smart library; deep learning; convolution neural network; TensorFlow; automatic classification

1 引言:图书馆与文献自动分类

文献的标引编目加工是图书馆重要的业务工作之一,其工作量大,专业性强,又是需要多人协作的综合性工作,有自己的特点和规律,主要采用手工分类的方式。在知识爆炸的时代,需要对数量庞大、内容复杂、形式多样的文献进行准确的归类、标引,对工作人员的要求很高;另一方面,由于编目外包和图书馆学专业教育的转型,资深标引编目人员日趋减少,信息加工质量和效率都呈下降趋势。

20世纪50、60年代在H.P.Luhn、Maron等人的推动下,图书馆界一直在探索文献自动分类的方法。国内相关研究起始于上世纪80年代初[1]。近年来随着人工智能技术的蓬勃发展,文献自动分类由基于规则的分类转向基于机器学习的分类,旨在提高文献的分准率。

1.1 基于规则的分类方法

基于规则的分类方法主要包括基于词典发的分类方法,即构建主题词与分类号的对照关系表,扫描并找出文章所包含的主题词进而计算文献的类归属

收稿日期:2017-12-21;责任编辑:魏志鹏

性;基于专家系统的自动分类方法,即构建专家系统结合推理机实现文献分类[1]。此类方法的一方面构建分类主题词表,但由于在知识爆炸的当下,各学科发展迅猛文献内容、形式多样使得词表的编制滞后于科学的发展,使得其对于包含新词的文献无法分类;另一方面经常需要人工依学科发展的情况不断调整分类规则。

1.2 基于机器学习的分类方法

文献分类过程实质是编目人员依据文献题名、关键词和摘要结合其对中图分类法的理解赋予一个中图分类号的过程(少数情况下需要通读全文)。换言之,即是编目人员通过培训学习中图分类法构建相应的分类体系,利用培训学习的成果对文献进行加工,并在实践中不断完善自己的分类体系。把上述过程泛化,利用已编目的文献构建题名、关键词和摘要的知识库,提取相应的特征数据进行学习,这便是基于机器学习的分类方法。

基于机器学习的分类方法其基本过程主要包括:构建语料库、文本建模、特征选择、特征扩展、选择并实现分类算法五个环节。常用的方法有朴素贝叶斯法、KNN、决策树法、中心向量法、支持向量机以及近两年兴起的人工神经网络的分类方法等。基于神经网络的分类方法虽在小规训练集上与其他传统的机器学习分类方法不相上下,但随着数据集和网络规模的增大,其性能远超于传统的机器学习方法,能够更好处理海量数据(见图1)。

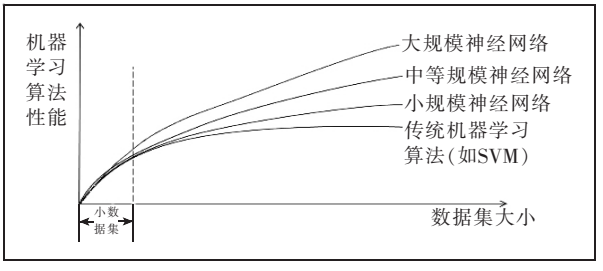

图1 数据集大小与机器学习算法性能比较

本文尝试将卷积神经网络引入到文献自动分类研究中,构建基于题名、关键词的多层次卷积神经网络网络模型,使之能够根据文献的题名和关键词自动给出中图分类号,以解决编目人员紧缺,加工质量和效率下降的问题,并在实际生产环境下证明该模型的准确性和合理性。

2 深度学习与TensorFlow

随着第三次人工智能浪潮的兴起,机器学习作为一种数据挖掘的方法被广泛应用于垃圾邮件检测、定向客户的产品推荐、商品预测等领域。近年来,受益于计算机在通用计算领域计算性能的持续提升和海量数据的便捷获取,深度学习作为一种特殊的机器学习范式在图像识别、语音识别、机器翻译、文本分类等领域获得巨大成功,凭借从输入数据中判断"哪些是特征值",无需人工干预的能力,其在医疗诊断、艺术创作、医疗诊断、自动驾驶等更加复杂的领域也有突破性的进展,并已开始应用于实际工作中。

2.1 深度神经网络

深度学习神经网络是人工神经网络的扩展,人工神经网络是基于模拟大脑皮层的神经网络结构和功能而提出的计算模型(见图2),人工神经元细胞可根据输入信号 p_i 的刺激触发输出a,大量的人工神经元细胞依一定的规则(即权重 w_i)连接在一起形成一个大规模并行计算网络,即人工神经网络。

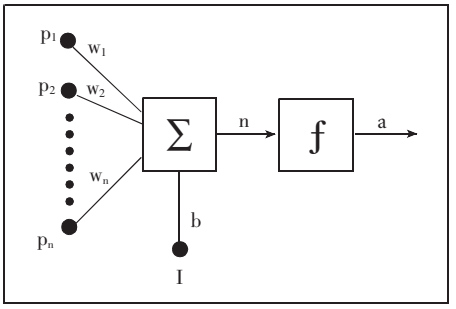

图2 人工神经元数学模型图

相较于其他机器学习方法,深度学习在模拟人脑神经元间的连接、对外界刺激的感知和传导的同时,采用让各层预先学习的方式,建立对观察数据(或称训练数据、输入)和标签(或称输出)之间的联合分布。学习从浅层顺次开始,上一层学习得出的数据会作为下一层的输入数据,由浅层的初级特征逐步学习到深层的高级特征。如在学习什么是狗时,第一层是一个轮廓、下一层是眼、鼻子的形状,在下一层是脸上的其他细节。以此类推,是一个从全局到局部再到细节特征的学习过程,每一层都在分段学习,学习过程中的错误也可以在每一层得到相应处理,这使得其具有自我学习和解决问题的能力,该模型

最早由多伦多大学的 Hitton 教授于 2006 年提出——一种名为深度置信网络(Deep Belief Net,DBN)[9],在 2012 年的 ImageNet 图像识别大赛中以低于第二名 10%的错误率而崭露头角[10],之后 LeCun、Mikolov 等人则提出卷积神经网络和循环神经网络，对深度学习进行优化和扩展。

2.2 深度学习框架

为了更好、更方便高效使用机器学习算法,通常需要一定的软件平台支持，如 Caffe、Theano、Torch、CNTK、Tensorflow 等。

Tensorflow 是谷歌于 2015 推出的一种供机器学习所使用的利用数据流图进行计算的库套件，遵循 Apache2.0 协议。相对于其他几个神经网计算框架而言,Tensorflow 属于其中的后起之秀，它支持多种机器学习常用的开发语言(如 C++、Python、Cuda),支持几乎所有类型的深度学习算法的开发（如 CNN、RNN、LSTM 等),能在多种硬件环境(CPU、GPU、TPU 手机、云)下很好地利用各自的长处和特点运行,并能够进行网络分布式学习。由于其具有众多优点,如计算速度快、部署容易、灵活性强、可扩展等,有学者在 github 上发布了关于 Caffe,Theano、Torch、CNTK、Tensorflow 性能比较的文章,从网络模型能力、接口、模型部署、性能,架构和跨平台方面对其进行比较分析并做相应评分（满分为 5 分)[11](见表 1)，比较可见,Tensorflow 无论是单项还是综合评分都比较高。

Google 是 TensorFlow 的最大用户和推动者,在谷歌的强力推广下,很多高校、科研机构和第三公司已开始使用 Tensorflow,例如谷歌利用该平台对其自动翻译服务进行了系统升级，翻译质量比过去有明显提升；在谷歌邮件系统中，用 sequence-to-sequence[12]模型来自动建立文本摘要,并对邮件语境预测可能的回复;对视网膜影像数据进行训练,已能成功预测影像是否有糖尿病引起的视网膜病变[13];在 AutoDraw[14]中开发“预测”功能,可以根据标题和用户画出的部分元素推测并继续完成一幅绘画作品; Google Now 则通过适当的数据反馈(RNN,反馈神经网络)来理解音频信号,进而实现语音识别、语音搜索、语音情感分析等[15]。这些科研应用也给深度学习在其他行业中的应用提供了参照。

表 1 各神经网络框架评价比较

软件平台	网络和模型能力	接口	模型部署	性能(单 GPU)	架构	跨平台
Caffe	3	3	5	–	3	√
CNTK	2	2.5	4.5	–	–	√
Tensorflow	4.5	4.5	4.5	5	5	√
Theano	4.5	4	3	3	3	√
Torch	5	4	3	5	5	不windows

3 基于卷积神经网络的《全国报刊索引》文献分类模型

《全国报刊索引》近 4 年历史数据约为 170 万条,包含题名、关键词、分类号、摘要、作者、出版社、全文等文献信息。一方面由于文献题名与内容有着较高的符合率[1]，且题名是一个有限长度、结构紧凑、能够表达独立意思的短句,这使得卷积神经网络可以用于文献的分类；另一方面从摘要中提取正确关键词存在一定困难,所以本文选取题名+关键词作为网络模型训练的训练集，文献对应的中图法分类号作为网络模型的输出。

3.1 文献分类系统模型设计

基于深度学习的报刊索引文本分类基本思想是将已分好类的文献题名和关键词经切词后构成二维词向量作为神经网络的输入,分类号作为输出,通过多层神经网络训练后,对新的文献分类进行预测。本文所用数据中,中图分类法一级类目 38 个、四级类目 9668 个,为了降低训练成本,本文模型采用粗、细分类的分层分类结构(见图 3),先大类分类,随后在大类分类

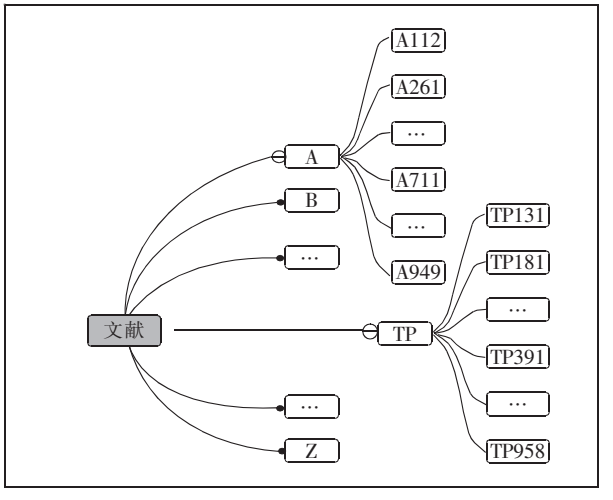

图 3 基于文献粗、细分类结构的层次分类结构

的基础上将其进行四级分类;预测也是如此。

分类系统采用模型预训练和模型预测组成。其中，预训练是通过对现有文献分类的结果搭建深层神经网络的深度学习模型并进行数据训练，包括数据预处理和机器学习两部分；模型预测则是对未知文献进行分类结果预测(见图 4)。

3.2 数据预处理

由于神经网络的准确率对于受训练数据影响较大，故数据预处理是整个系统的第一步也是最为关键的一步,包含分词、词向量转换以及输出标签的独立热编码(one-hot code)。

3.2.1 分词

分词则是将自然语言转换为一组词语的表达,与英文依空格切词不同,中文分词分为句子切分,对输入的中文文档进行预处理，得到单个中文短句的集合;原子切分,对输入的中文短句进行原子切分,并根据所得的原子系列建立初始的切分词图；堆砌词语,基于原子系列,从不同视角分别进行中文词语识别，并将各自的堆砌结果添加到切分图；分词优选,基于上一阶段的堆砌路径和各路径的概率,计算得出最可能的堆砌路径,作为最后的分词结果,并输出最终结果,四个步骤。本文采取的做法如下:首先对所有文献的关键词做词频统计，并构建分词用主题词表;基于前缀词典实现高效的词图扫描,结合主题词表生成句子中汉字所有可能成词情况所构成的有向无环图；其次运用动态规划算法查找最大概率路径,并找出基于词频的最大切分组合;对于未登录词，采用隐马尔可夫模型（Hidden Markov Model，HMM)[24]模型做汉字成词处理。

3.2.2 词向量

正如前文所提到的,深度学习实质是数值计算,所以需要词向量转换将自然语言转换成可计算的数学表达，即将一个词转换成一定空间向量下的概率表达即 p=(w(t)｜(w(t-n+1),…,w(t-1)),其中 w(t) 为句子中第 t 个词在文本中的向量表达。word vector 则表示由该文献题名和关键词组成的词向量组(见图 5)。

$$word\ vector=\begin{bmatrix}\begin{bmatrix}wv_{11} & wv_{12} & \cdots & wv_{1j} & \cdots & wv_{1m}\\ wv_{21} & wv_{22} & \cdots & wv_{2j} & \cdots & wv_{2m}\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ wv_{i1} & wv_{i2} & \cdots & wv_{ij} & \cdots & wv_{im}\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ wv_{n1} & wv_{n2} & \cdots & wv_{nj} & \cdots & wv_{nm}\end{bmatrix} label\end{bmatrix}$$

图 5 文献的词向量矩阵

其中,label 表示文献所对应的分类号采用独立热编码形式,将分类号映射为 N 维空间向量(N 为总分类个数),当某一个维度上的值为 1,其它位为 0 时表示该表示其所对应的分类号，即 label=((1 0 … 0 0));词向量 $wv=(wv_{11}\ wv_{12}\ \cdots\ wv_{1j}\ \cdots\ wv_{1m})$表示该文献的中一个词。

词向量分为静态(static)和非静态(non-static)方式两种,静态方式采用预训练的词向量,训练过程不更新词向量，在数据量不大的情况下使用静态方式可以得到不错的效果；非静态方式则是在训练过程

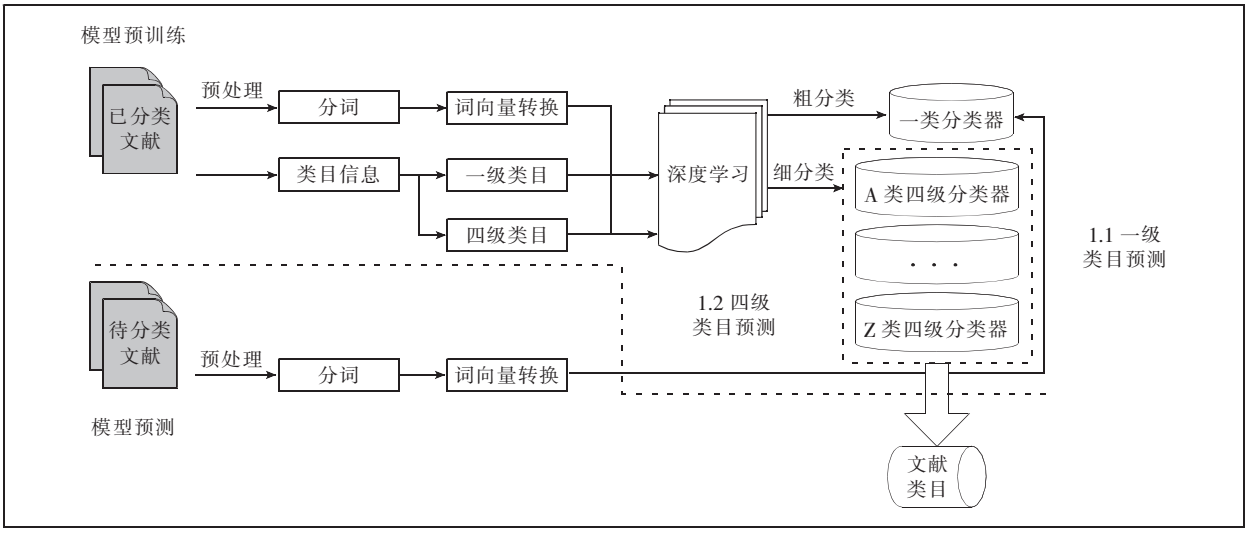

图 4 基于深度学习的文献自动分类系统

中更新词向量，训练过程中调整词向量，能加速收敛。词向量训练模型有很多如 skip-gram、CBOW[20-22]、C&W[23]模型等，本文采用静态方式，使用 skip-gram 模型，结合文献的题名、关键词和摘要的分词结果作为词向量的训练集，构建静态词向量。

3.3 卷积神经网络分类模型的分析与设计

通过 Yoon kim 的研究表明，有限长度、结构紧凑、能够表达独立意思的句子可以使用卷积神经网络进行分类[18,25,26]，在其研究的基础上，本文提出将文献的题名、关键词作为训练集，并搭建多层卷积神经网络用于文献分类的训练和预测。

卷积神经网络（convolutional neural network, CNN）[27]顾名思义，将卷积滤波与神经网络两个思想结合起来与普通神经网络的区别在于，卷积神经网络包含了一个由卷积层和子采样层构成的特征抽取器。在卷积神经网络的卷积层中，一个神经元只与部分邻层神经元连接。在 CNN 的一个卷积层中，通常包含若干个特征平面(feature maps)，每个特征平面由一些矩形排列的的神经元组成，同一特征平面的神经元共享权值，这里共享的权值就是卷积核。卷积核一般以随机小数矩阵的形式初始化，在网络的训练过程中卷积核将学习得到合理的权值。共享权值(卷积核)带来的直接好处是减少网络各层之间的连接，同时又降低了过拟合的风险。子采样也叫做池化(pooling)，通常有均值子采样(average pooling)和最大值子采样(max pooling)两种形式。子采样可以看作一种特殊的卷积过程。卷积和子采样大大简化了模型复杂度，减少了模型的参数。为此，我们可以得知卷积神经网络的基本结构(见图 6)，其由三部分构成。第一部分是输入层；第二部分由 n 个卷积层和池化层的组合组成；第三部分由一个全连结的多层感知分类器构成。

在作者文献自动分类的神经网络结构中（见图 7)，输入层为 20*20 词向量，隐含层由卷积核为 2*20 和 3 个 2*1 的卷基层堆叠而成，输出层为全连接层，结合 softmax 激活函数将提取的文本特征输出为各个分类上的概率分布。

3.4 实验结果与分析

实验所用数据为上海图书馆《全国报刊索引》2013-2016(或 2014-2017 年，作者确定年份)年 170 万余条题录将上述模型于 TensorFlow 平台上进行训练和调试，其中训练集为 153 万条，训练用验证集为

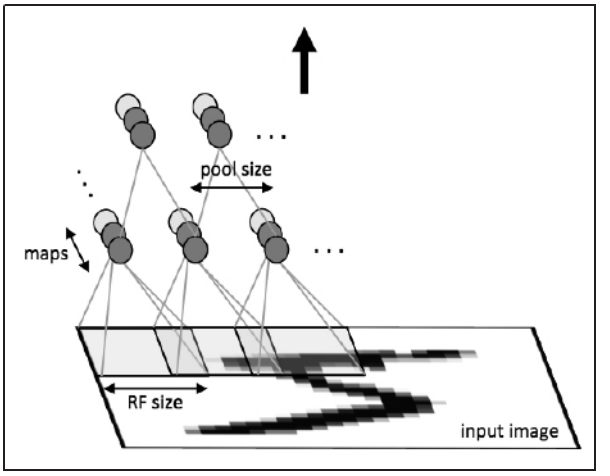

图 6 卷积神经网络基本结构

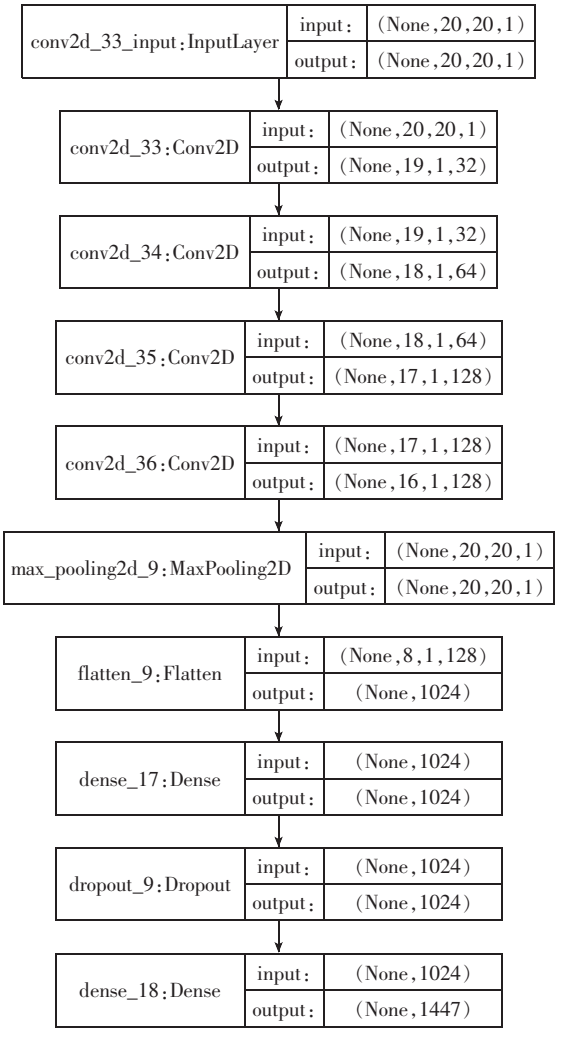

图 7 文献自动分类神经网络结构

17万条。模型训练集的准确率收敛于67%,训练用验证集的准确率收敛于69%左右(见图8、9)。

在生产环境中，模型预测结果的正确与否是以人工分类结果为参照标准。对未知的7144条待加工数据做分级分类预测,并与人工分类结果做比较,测试后得知，一级准确率为75.39%，四级准确率为57.61%(见表2)。

正如前文所提到神经网络的结果输出是为一个分类上的概率表达,当设输出阈值(置信度)为0.9时，虽模型一级正确率为43.98%，一级错误率为1.96%,四级输出正确率为25.66%,错误率为5.11%(见表3)。这表明对于测试集而言其预测结果在阈值为0.9时的输出结果有着较低的错误率,即拥有较高的可信度。

3.4.1 训练集对准确率的影响

受期刊收录稿件偏好影响，本文所使用的数据存在很大的不均衡性(见表4),大量的数据集中在D、F、G、R四个大类上，最少的Z大类只有20个训

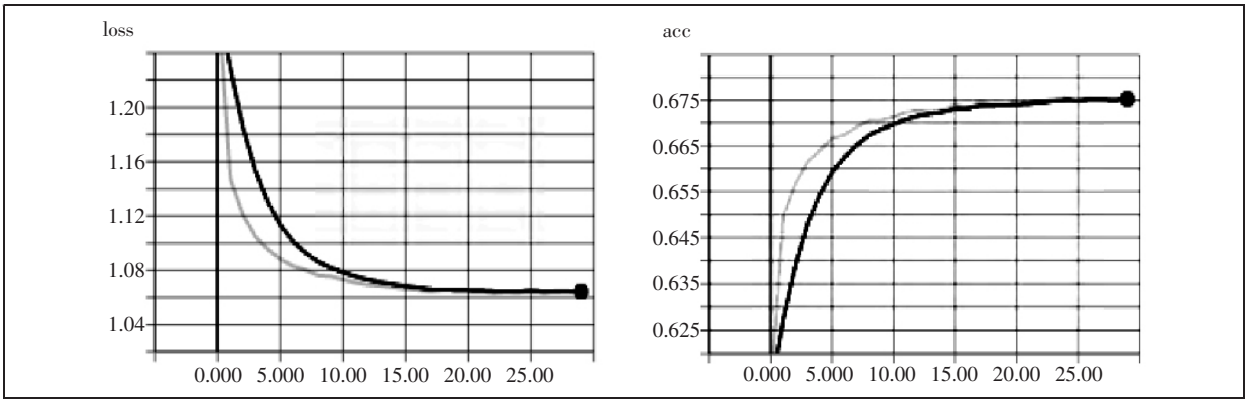

图8 模型训练集的损失函数及准确率函数

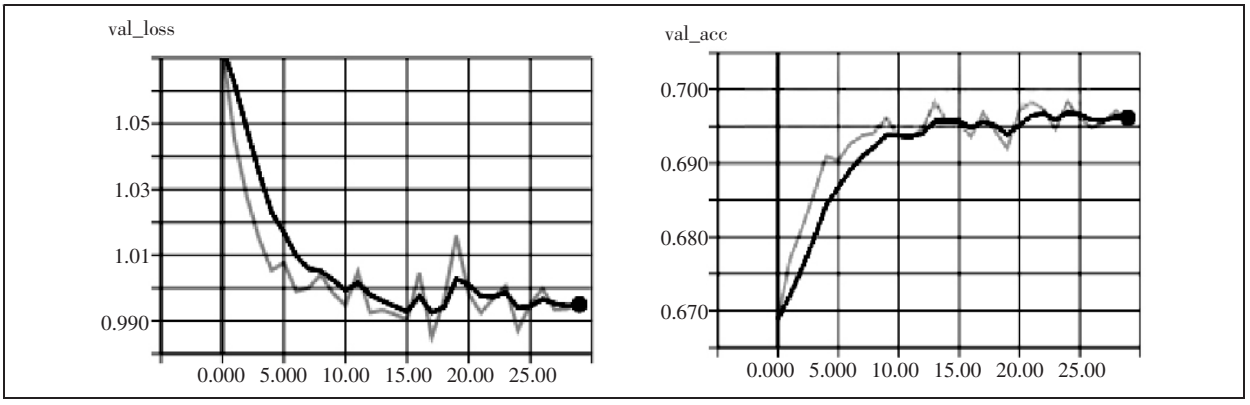

图9 模型测试集的损失函数及准确率函数

表2 验证测试结果

数据量	总计	一级正确	占比	二级正确	占比	三级正确	占比	四级正确	占比
社科	3027	2368	78.23%	2338	77.24%	2304	76.11%	2304	76.11%
科技	4117	3018	73.31%	2376	57.71%	2120	51.49%	1812	44.01%
总计	7144	5386	75.39%	4714	65.99%	4424	61.93%	4116	57.61%

表3 阈值为0.9时的验证测试结果

数据量	总计	一级正确	正确率	一级错误	错误率	四级正确	正确率	四级错误	错误率
社科	3027	1267	41.86%	54	1.78%	1071	35.38%	51	1.68%
科技	4117	1875	45.54%	86	2.09%	762	18.51%	314	7.63%
总计	7144	3142	43.98%	140	1.96%	1833	25.66%	365	5.11%

表 4　训练用各大类样本数

类别	样本数	类别	样本数	类别	样本数	类别	样本数
A	4662	K	26971	TE	9694	TQ	26149
B	23804	N	1329	TF	4827	TS	24214
C	10412	O	30542	TG	29013	TU	21466
D	116740	P	34934	TH	10215	TV	4966
E	8738	Q	18020	TJ	3358	U	22060
F	162809	R	196364	TK	6546	V	12503
G	137052	S	55494	TL	3736	X	33685
H	23447	T	64	TM	29155	Z	20
I	37892	TB	17518	TN	32858		
J	28643	TD	8048	TP	49312		

练样本,由于神经网络的训练集不均衡性[28-31]导致模型预测准确率在一定层度下会有所下降，通过训练集均衡以获得最佳结果；对于一些极度不均衡的数据，如T、Z大类，四年内总计数据不足100条的类目,无法做样本均衡,则可将其标统一注为“未知”类目,当预为“未知”类目时,直接交由人工处理。

3.4.2 分词对准确率的影响

由于中文分词的特殊性，使得分词的分准率受词表影响较大,如“上海图书馆”一词,在没有相应的主题词表时会被切割为“上海/图书馆”,使得其在句中的意思是有所改变,影响训练时的特征提取,进而影响预测的准确性。由于并无主题词表,故对170万分文献中出现的关键词做词频统计，为分词提供主题词表。经不完全测试,在有无词表的情况下,准确率相差约2%。

3.4.3 验证数据对结果的影响

由于验证数据采用实际生产环境中的数据作为测试集,其数据集合并不覆盖所有的分类项目,且存在数据不均衡问题,使得测试结果不具有统计学,但反映了其在实际生产环境下的使用情况，证明基于卷积神网络的文献自动分类在实际工作中的可行性。

4　展望

谷歌的最新研究成果表明，将计算机视觉和语言模型通过CNN与RNN网络叠加进行合并训练,所得到的系统可以自动生成一定长度的文字文本[19]等。这些研究成果非常适合应用于图书馆内部业务的智能化上，如图书馆藏资源的自动分类、自动摘要、主题提取、文章聚类、图片自动标引、图像识别、业务预测和分析等。

本文在对深度学习的研究基础上提出了基于深度学习的文献自动分类模型，将文献分类问题转化为基于神经网络的自动学习和预测的问题。通过对《全国报刊索引》170万条数据的模型训练以及7000多篇待加工的文献预测,证明此方法是可行的,且有较高的置信度,分词、词表、模型训练完全依赖于历史数据但本文仅细分至四级类目,随着分类的逐步深入,题名与关键词并不能很好的体现出文献之间的差异。摘要是体现文献细微差别的重要切入点,在接下来的研究中，将会研究如何从摘要中提文献信息,以提升分类准确率和细分程度。

参考文献：

[1] 成颖,史九林.自动分类研究现状与展望[J].情报学报,1999,18(1):20-26.

[2] 李湘东,阮涛,刘康.基于维基百科的多种类型文献自动分类研究[J/OL].[2017-10-17].http://kns.cnki.net/kcms/detail/11.2856.G2.20171017.1501.012.html.

[3] 张野,杨建林.基于KNN和SVM的中文文本自动分类研究[J].情报科学,2011,29(9):1313-1317.

[4] Wei L,Wei B,Wang B,et al.Text Classification Using Support Vector Machine with Mixture of Kernel[J].Journal of Software Engineering and Applications,2012,5(12):55-58.

[5] Hebb Donald.The Organization of Behavior a neuropsychological theory[M].New York:John Wiley,1949:100-136.

[6] Liu M Q.Discrete-time delayed standard neural.Network and its application[J].Sci China,2006,49(2):137-154.

[7] 王昊,严明,苏新宁.基于机器学习的中文书目自动分类研究[J].中国图书馆学报,2010,36(6):28-39.

[8] 叶鹏.基于机器学习的中文期刊论文自动分类研究[D].南京:南京大学,2013.

[9] Hinton G E,Osindero S,Teh Y W.A fast learning algorithm for deep belief nets[J].Neural computation,2006,18(7):1527-1554.

[10] Krizhevsky A,Sutskever I,Hinton G E.ImageNet classification with deep convolutional neural networks[C].International Conference on Neural Information Processing Systems.Curran Associates Inc,2012:1097-1105.

[11] Evaluation of Deep Learning Toolkits[EB/OL].[2017-10-17].https://github.com/zer0n/deepframeworks/blob/master/README.md.

[12] Ilya Sutskever,Oriol Vinyals,Quoc V.Le.Sequence to sequence learning with neural networks[C].Advances in Neural Information Processing Systems 2014:3104-3112.

[13] google developers blog[EB/OL].[2017-10-17].https://developers.googleblog.com/2017/02/announcing-tensorflow-10.html.

[14] Auto Draw[EB/OL].[2017-10-17].https://www.autodraw.com/.

[15] Alex Graves,Abdel-rahman Mohamed,Geoffrey Hinton. Speech recognition with deep recurrent neural networks[C].International Conference on Acoustics,Speech and Signal Processing,2013:6645-6649.

[16] Maron M E.On Relevance,Probabilistic Indexing and Information Retrieval[J].Journal of the Acm,1960,7(3):216-244.

[17] 刘佳宾,陈超,邵正荣,等.基于机器学习的科技文摘关键词自动提取方法[J].计算机工程与应用,2007(14):170-172.

[18] Yoon Kim.Convolutional Neural Networks for Sentence Classification[C].Empirical Methods in Natural Language Processing(EMNLP),2014:1746-1751.

[19] A Picture is Worth Thousand Coherent[EB/OL].[2017-10-17].https://research.googleblog.com/2014/11/a-picture-is-worth-thousand-coherent.html.

[20] Tomas Mikolov,Kai Chen,Greg Corrado,et al.Efficient Estimation of Word Representations in Vector Space[J].arXiv preprint,2013: arXiv:1301.3781.

[21] Tomas Mikolov,Ilya Sutskever,Kai Chen,et al.Distributed Representations of Words and Phrases and their Compositionality[C].Advances in Neural Information Processing Systems,2013:3111-3119.

[22] Yoav Goldberg,Omer Levy.word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method[J].arXiv preprint,2014:arXiv:1402.3722.

[23] Collobert R,Weston J.A unified architecture for natural language processing:deep neural networks with multitask learning[C].International Conference.DBLP,2008:160-167.

[24] Kevin P.Murphy,Mark A.Paskin.Linear Time Inference in Hierarchical HMMs[C].Proceedings of Neural Information Processing Systems,2001:833-840.

[25] Nal Kalchbrenner,Edward Grefenstette,Phil Blunsom.A convolutional neural network for modelling sentences[J].arXiv preprint.2014: arXiv:1404.2188.

[26] Ying Wen,Weinan Zhang,Rui Luo,et al.Learning text representation using recurrent convolutional neural network with highway layers[J].arXiv preprint,2016: arXiv:1606.06905.

[27] LeCun,Yann.LeNet-5,convolutional neural networks[EB/OL].[2017-10-17].http://yann.lecun.com/exdb/lenet/.

[28] Paulina Hensman,David Masko.The impact of imbalanced training data for convolutional neural networks[EB/OL].[2017-10-17].https://www.kth.se/social/files/588617ebf2765401cfcc478c/PHensmanDMasko_dkand15.pdf.

[29] Palodeto V,Terenzi H,Marques J L B.Training neural networks for protein secondary structure prediction: the effects of imbalanced data set[C].Intelligent Computing,International Conference on Emerging Intelligent Computing Technology and Applications.Springer-Verlag,2009:258-265.

[30] Chandonia J M,Karplus M.The importance of larger data sets for protein secondary structure prediction with neural networks.[J].Protein Science,2010,5(4):768-774.

[31] Pulgar F J,Rivera A J,Charte F,et al.On the Impact of Imbalanced Data in?Convolutional Neural Networks Performance[C].International Conference on Hybrid Artificial Intelligence Systems.Springer,Cham,2017:220-232.

作者简介:郭利敏,男,上海图书馆系统网络部工程师。