基于卷积神经网络的中文新闻文本分类

蓝雯飞,徐 蔚,王 涛
(中南民族大学 计算机科学学院,武汉 430074)

摘 要 经典的卷积神经网络文本分类模型仅仅着眼于全局特征,没有考虑到局部特征.为了解决此问题,引入了注意力机制,用于提取文本中的关键词,把全局特征与局部特征综合在一起,使得文本的特征表达更加丰富.实验结果表明:卷积神经网络分类模型比传统的机器学习方法分类效果更好,而引入注意力机制后的卷积神经网络模型相比于经典的文本分类模型,分类效果也有了一定程度的提高.

关键词 自然语言处理;深度学习;卷积神经网络;注意力机制;文本分类
中图分类号 TP183 **文献标识码** A **文章编号** 1672-4321(2018)01-0138-06

Text Classification of Chinese News Based on Convolutional Neural Network

Lan Wenfei, *Xu Wei*, *Wang Tao*
(College of Computer Science, South-Central University for Nationalities, Wuhan 430074, China)

Abstract The classical convolutional neural network text classification model only focuses on the global features, without taking into account the local features. To solve this problem, the attention mechanism is introduced to extract keywords from the text. In this way, the global features and local features are combined together, which makes the feature representation of the text richer. Experimental results show that the text categorization model of Convolutional Neural Network is better than the traditional machine learning methods. Compared with the classical text classification model, after introducing attention mechanism, the performance of Convolutional Neural Network classification model has been improved.

Keywords natural language processing; deep learning; Convolutional Neural Network; attention mechanism; text classification

在众多种类的数据信息中,中文新闻文本数据是数据信息的一种重要承载形式,文本分类技术作为一种高效的信息检索与数据挖掘信息技术,在对文本数据信息的组织与管理中具有举足轻重的地位.

文本分类作为分类研究的一个重要研究方向,其分类方法与其他的分类没有本质的区别,核心方法为提取分类数据的特征和选择最优的特征匹配.但是文本也有自己的特点,根据文本的特点,文本分类的流程为:预处理、文本表示以及特征选择、构造分类器.预处理阶段主要是进行分词和去停用词,因为中文文本不像英文文本那样是以空格隔开,一个单词即为一个词,所以要利用分词算法针对中文文本进行分词处理.其次,要去掉常用且意义不大的停用词,如:“的”、“了”、“是”等.传统的文本表示方法是基于词袋模型(BOW)[1],该模型只包含词的词频信息,而忽略了句子的上下文关系的表示形式.这种浅层的文本表示对原始文档的语义表达造成了极大的信息损失,使得后续的文本挖掘任务面临巨大的挑战.本文以 word2vec 训练得到的中文词向量为表示方法,该方法克服了传统词袋模型的缺点,能够体现词之间的语义相关性.对于特征选择,常用的特征计算有 TF-IDF,互信息量,信息增益,χ^2 统计量等方法,然后利用选取的特征构造分类器进行分类,传统的机器学习分类方法有:最近邻(KNN)、朴素贝叶斯(NB)以及支持向量机(SVM)等.

目前,基于深度模型,从大规模的语料中挖掘文本的深度语义信息已经成为自然语言处理领域的热

收稿日期 2017-12-06
作者简介 蓝雯飞(1966-),女,教授,研究方向:数据库,软件技术,E-mail: lanwenfei1@163.com
基金项目 国家自然科学基金资助项目(61379059)

门研究方向.深度学习模型展示出了前所未有的表达能力,具有较强的理论意义.从实际应用角度出发,伴随着互联网规模的扩张和多媒体的发展,大规模训练数据以及机器设备性能的提升,高性能 GPU 集群提供了强大的计算能力,给深度学习提供了一个技术革新的舞台.本文使用的卷积神经网络(CNN)是一种深度监督学习下的机器学习模型,具有较强的适应性,善于挖掘数据的局部特征.与传统机器学习需要手动提取特征的方法相比,CNN 能够自动提取全局特征,它的权值共享结构网络使之更类似于生物神经网络,在模式识别各个领域都取得了较好的成果.

1　相关研究工作

Collobert 等[2]提出了使用神经网络的方法自动学习词汇的向量化表示,其基本原则是:一个词包含的意义应该由该词周围的词决定.首先将词汇表中的每一个词随机初始化为一个向量,然后用大规模的语料作为训练数据来优化此向量,使相似的词具有相近的向量表示.这样的训练方法能够将适合出现在窗口中间位置的词聚合在一起,而将不适合出现在这个位置的词分离开来,从而将语义(语法或者词性)相似的词映射到向量空间中相近的位置.与替换中间词的方法不同,Mikolov 等[3,4]提出了一种使用周围词预测中间词的连续词袋模型(CBOW).连续词袋模型将相邻的词向量直接相加得到隐层,并用隐层预测中间词的概率.同词袋模型一样采用的是直接相加,所以周围词的位置并不影响预测的结果.Mikolov 等人还提出了一种连续 Skip-gram 模型.同连续词袋模型的预测方式相反,连续 Skip-gram 模型通过中间词来预测周围词的概率.

Kim[5]在不同的分类数据集上评估卷积神经网络模型,主要是基于语义分析和话题分类任务.输入层是一个表示句子的矩阵,每一行是 word2vec 词向量.接着是由若干个滤波器组成的卷积层,然后是最大池化层,最后是 softmax 分类器.Zhang 等[6]通过多次重复实验,比较了不同超参数对 CNN 模型结构在性能和稳定性方面的影响.

本文针对中文新闻文本,通过 word2vec 训练大量的中文新闻语料得到中文词向量.从训练好的词向量中找出分类文本对应的词向量,作为文本分类的输入特征.然后构建 CNN 神经网络模型,调整不同的超参数,并在模型中加入了注意力机制,最终提高了文本的分类效果.

2　相关理论

2.1　文本特征表示

在自然语言处理(NLP)任务中,我们让机器学习算法,来处理自然语言,但机器无法直接理解人类的语言,所以要将语言数学化.如何对自然语言数学化,词向量就提供了一种很好的方式.

一种最简单的词向量是 One-hot Representation,就是用一个很长的向量来表示一个词.其中向量的长度为词典 D 的大小 N,向量的分量只有一个“1”,其他的全为“0”,“1”的位置对应该词在词典中的索引.这种表示方法非常简单,但是也存在一个重要的问题——“词汇鸿沟”现象:任意两个词之间都是孤立的,仅仅从这两个向量中看不出两个词是否有关系.另外,词典中的词汇量肯定是很大的,这种表示方法会使向量的维数很大,易发生维数灾难.

另一种表示方法是 Distributed Representation,它最早是 Hinton[7,8]于 1986 年提出的,可以克服 One-hot Representation 的缺点.其基本思想是:通过训练将某种语言中的每一个词映射成一个固定长度的短向量,将所有这些向量放在一起形成一个词向量空间,而每一向量为该空间中的一个点,在这个空间引入“距离”,则可以根据词之间的距离来判断它们之间(词法、语义上)的相似性了.

2.2　NLP 中的 CNN 模型

CNN 卷积神经网络是 20 世纪 60 年代 Hubel 和 Wiesel[9]提出来的,如今 CNN 成为了深度学习研究的热点之一,在图像处理领域应用广泛[10-13].CNN 其实是由多层的神经网络组成的,除了输入层和输出层以外,还包括特征提取层以及特征映射层.与传统的神经网络不同,CNN 有两个特点:(1)局部感知;(2)权值共享.相比于传统的神经网络,这两个特点大大减少了神经网络训练过程中的参数数量.

在图像处理任务中,CNN 输入是以像素点为单位进行一系列操作运算的,而在 NLP 任务中,是以词向量为单位.这里的词向量表示方法可以是 One-hot 的形式,也可以是由 word2vec 训练得到的 Distributed Representation 的形式.图 1 所示的是用于文本分类的基本 CNN 模型,也是 Kim 在论文中提到的模型,其中的数字标识为 Kim 论文中的主要相关参数设置.

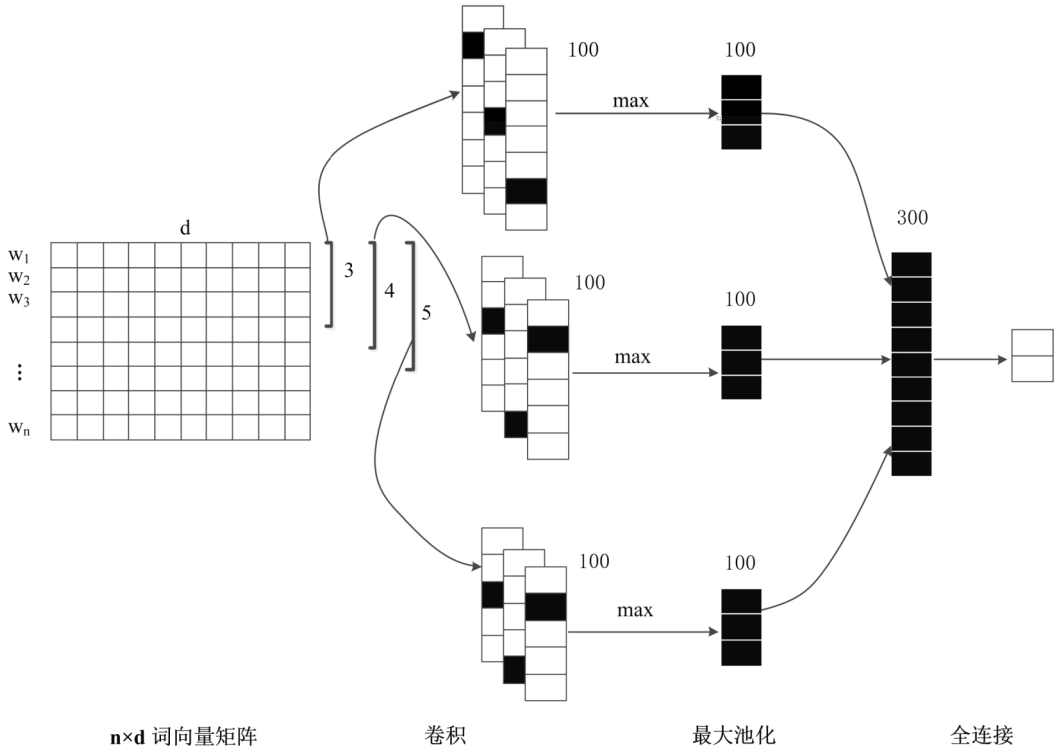

图1 NLP中的CNN模型

Fig.1 CNN model in NLP

3 基于注意力机制的CNN模型

注意力机制最早是在图像领域提出来的[14],其研究是受人类注意力机制的启发.而在NLP领域,最早是在机器翻译上应用[15,16].机器翻译使用的是一种典型的序列到序列(sequence to sequence)的模型,也是一种编码到解码(encoder to decoder)的模型.传统的机器翻译模型仅根据最后一个词学到的表达和当前要预测翻译的词联系起来,而加入注意力机制后,将源语言端的每个词学到的表达和当前要预测翻译的词进行联系.相比传统的机器翻译,加入注意力机制后,效果有明显的提高.图2所示的是本文使用的基于注意力机制的CNN模型.该模型分为两个部分:一部分是全局特征提取,另一部分是局部特征提取.

3.1 全局特征提取

全局特征提取部分与2.2节提到的NLP中的CNN模型一样,最左边的输入层是一个$n \times d$的二维矩阵,其中n代表一个由词$w_1, w_2, w_3, \cdots, w_n$组成的句子的长度,$d$代表每个词的词向量的维度.模型特征提取部分主要包括卷积、池化等操作.

与图像领域特征提取一样,利用卷积运算来提取文本全局特征.在NLP中,通常对几个词的词向量进行卷积运算,图2中矩阵的卷积运算是通过卷积核从上到下滑动进行的.本文卷积核设计了3种尺寸,分别为$3 \times d$、$4 \times d$和$5 \times d$,每种尺寸的卷积核有128个,即可以学习128种特征.CNN的卷积运算如下:

$$c_i = f\left(\sum W_1 \cdot X_{i:i+h-1} + b_1\right), \tag{1}$$

其中c_i表示卷积运算后的结果,即输入矩阵与卷积核的点乘和加上偏置之后的激活输出.h为窗口大小,$X_{i:i+h-1}$为输入的第i个到第$i+h-1$个窗口内的词向量矩阵,W_1为卷积核即权值矩阵,b_1为偏置量,f为relu激活函数.

在通过卷积获得了特征之后,为了简化网络的计算复杂度,提取主要特征,利用池化操作对特征进行压缩.池化操作一般有两种:平均池化(Average Pooling)和最大池化(Max Pooling).本文采用的是最大池化,也就是选取每个特征中的最大值.将池化得到的所有最大值拼接起来,得到一个维度为384×1的特征向量.池化操作如下:

$$\hat{c} = \max\{c_1, c_2, c_3, \cdots, c_{n-h+1}\}, \tag{2}$$

其中$\hat{c}$表示最大池化运算后的结果,$c_i(i = 1, 2, \cdots, n-h+1)$为卷积运算后的结果.

3.2 局部特征提取

局部特征提取部分利用注意力机制提取出与分类类别紧密相关的关键词,输入层仍然是词向量矩阵.和卷积操作原理类似,设计了一个滑动窗口,窗口大小为k,与卷积操作不同的是每个窗口的权值是不

共享的.为了能够覆盖到所有的词,在输入矩阵首尾各加入$(k-1)/2$个随机初始向量,这样能保证窗口中的中心词都是原始向量矩阵中的词.

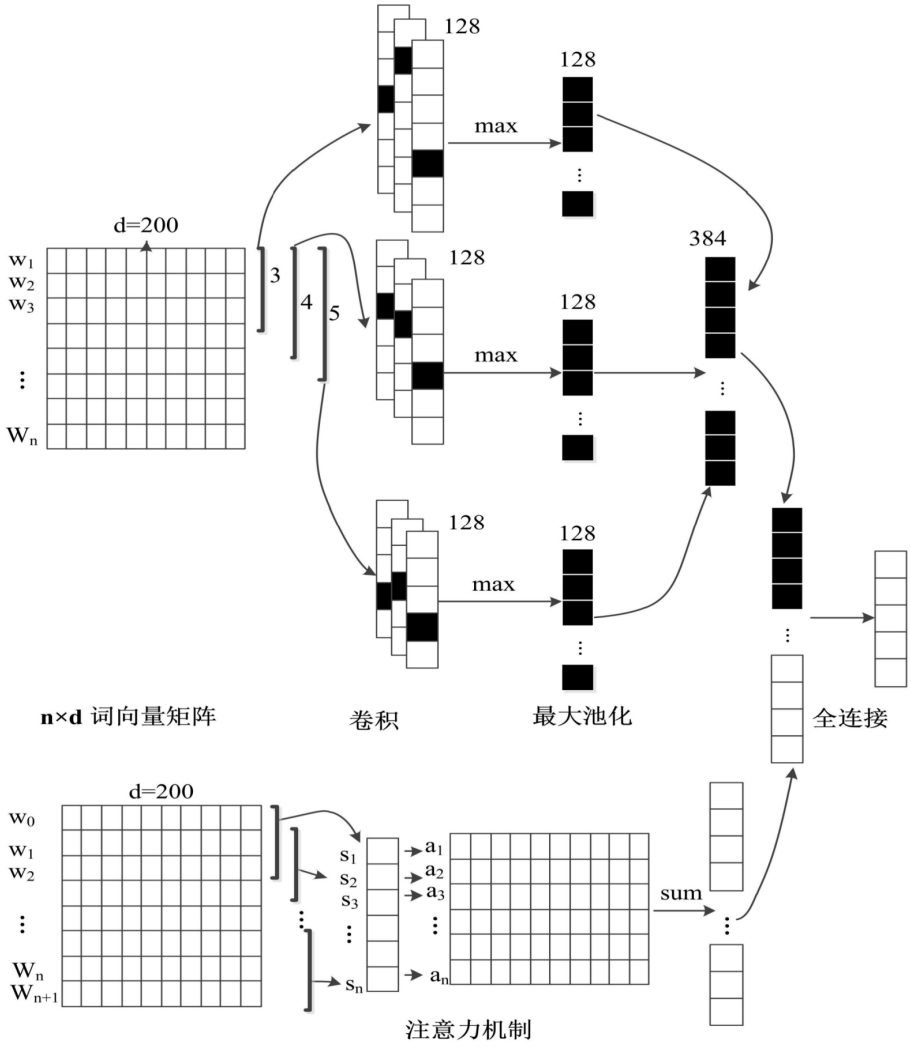

图 2　基于注意力机制的 CNN 模型

Fig.2　Attention-CNN model

为了评价各个词的重要程度,即判断某个词是否是关键词,有如下定义:

$$s_i = f\left(\sum W_2 \cdot X_{i:i+k-1} + b_2\right), \tag{3}$$

其中s_i为窗口中心词的权重,$X_{i:i+k-1}$为输入的第i个到第$i+k-1$个窗口内的词向量矩阵,W_2为窗口中词的权值矩阵,b_2为偏置量,f为 relu 激活函数.为了选择出关键词,定义一个关键词选择阈值λ,其定义如下:

$$\lambda = \frac{1}{n}\sum_{i=1}^{n} s_i. \tag{4}$$

通过阈值λ,选择大于阈值的关键词,得到关键词的向量a_i,其定义如下:

$$a_i = \begin{cases} w_i, s_i > \lambda, \\ 0, s_i \leqslant \lambda, \end{cases} \tag{5}$$

其中w_i为原始输入词的向量.

然后对关键词向量进行相加求和得到$\hat{w}$,定义如下:

$$\hat{w} = \sum_{i=1}^{n} a_i. \tag{6}$$

由关键词加和后的向量,得到输出特征向量$z \in \mathbb{R}^{r\times 1}$($r$为输出特征向量的维度,本文设置为 384),定义如下:

$$z = f\left(W_3 \cdot \hat{w}^{\mathrm{T}} + b_3\right), \tag{7}$$

其中$W_3 \in \mathbb{R}^{r\times d}$是权值矩阵,$b_3$为偏置量,$f$为 relu 激活函数.

利用卷积运算操作得到结果h_i,即全局特征向量与局部特征向量连接后的向量与卷积核的点乘和,然后加上偏置之后的激活输出,定义如下:

$$h_i = f\left(\sum W_4 \cdot [x \oplus z]_{i:i+r-1} + b_4\right), \tag{8}$$

其中 $W_4 \in \mathbb{R}^{r\times 1}$ 是卷积核即权值矩阵,b_4 为偏置量,$(\hat{c} \oplus z) \in \mathbb{R}^{2r\times 1}$ 为全局特征向量与局部特征向量连接操作,f 为 relu 激活函数.

最后将全局特征向量与局部特征向量连接后的向量,与所要进行分类的类别神经元进行全连接操作,得到预测类别为 y_i,定义如下:

$$y_i = softmax\left(\sum_j W_5 h_j + b_5\right), \tag{9}$$

其中 $W_5 \in \mathbb{R}^{c\times r}$($c$ 为类别数),b_5 为偏置量.

模型使用的代价函数定义如下:

$$H_{y'}(y) = -\sum_i y'_i \lg y_i, \tag{10}$$

其中 y'_i 是实际类别标签值,y_i 是利用 $softmax$ 激活函数计算的预测类别标签值.为了最小化该代价函数,模型训练采用梯度下降法.为了加快收敛以及减少计算量,对小批量样本采用梯度下降,即每次更新变量只需取一小批样本参加计算,本文实验中样本数设置为 128.另外,为了防止过拟合,在全连接层引入 Dropout[17] 策略.

4 实验

4.1 实验准备

本文的实验是在 Window10 系统下进行的,使用的 CPU 是 Inter Core i5-2450M 2.5GHz,内存大小为 6GB.实验编程语言为 Python3.0,开发工具为 Pycharm,使用到的深度学习框架为 Tensorflow1.0.1.本文的实验数据集来源于搜狗实验室中的搜狐新闻数据,从中提取出用于训练中文词向量的中文语料,大小约为 4GB.然后选出用于分类的中文新闻,本文实验提取了 5 个类别的新闻数据,分别为财经、汽车、娱乐、军事和体育.每个类别新闻为 2000 条,共 10000 条,利用这 10000 条数据进行十折交叉验证来评估模型的分类效果.

4.2 实验设计

本文采用 Attention-CNN 模型,对中文新闻文本进行分类.为了评价分类模型的效果,通过精确率(Precision)和召回率(Recall)以及 F_1 值对分类结果进行衡量.

(1)为了说明注意力机制对分类结果的影响,将本文使用的 Attention-CNN 模型的分类结果,与经典 CNN 模型分类结果进行对比实验.另外,为了说明基于 CNN 模型分类的优势,在同样的数据集上利用传统的机器学习分类模型对文本进行分类,然后将分类结果进行对比,使用的机器学习方法包括支持向量机[18]、最近邻[19]、朴素贝叶斯[20].为了排除由于特征构建方式的不同而导致实验结果没有可比性,利用传统机器学习的特征构建方式同样是基于词向量,每条新闻文本的特征取为所有词向量均值.

(2)为了说明 Attention-CNN 模型中的窗口大小 k 对提取局部关键词的影响,设计了不同的 k 值.利用 Attention-CNN 模型,对比了不同 k 值情况下分类的效果.

4.3 实验结果及分析

(1)表 1 所示的是不同分类模型的整体平均 Precision 值、Recall 值和 F_1 值的比较结果.

表 1 不同分类模型的平均分类结果比较

Tab. 1 Comparison of average classification results of different classification models

分类模型	Precision	Recall	F_1
NB	0.8800	0.8690	0.8698
KNN	0.9594	0.9590	0.9590
SVM	0.9653	0.9654	0.9653
CNN	0.9751	0.9750	0.9750
Attention-CNN-5	0.9830	0.9830	0.9830

(2)表 2 所示的是注意力机制中不同窗口大小 k 对整体平均 Precision 值、Recall 值和 F_1 值的影响.

表 2 不同 k 值的平均分类结果比较

Tab.2 Comparison of average classification results of different k values

分类模型	Precision	Recall	F_1
Attention-CNN-1	0.9733	0.9730	0.9730
Attention-CNN-3	0.9771	0.9770	0.9770
Attention-CNN-5	0.9830	0.9830	0.9830
Attention-CNN-7	0.9751	0.9750	0.9750

从表 1 可以看出,在相同的数据集上,以 word2vec 训练的词向量作为文本特征,除了 NB 分类模型分类效果较差,其他分类模型均取得了较好的分类效果,可以证明 word2vec 训练的词向量能够很好地描述文本特征.另外,CNN 分类模型以及 Attention-CNN 分类模型的分类效果要比传统机器学习分类模型好,原因在于 CNN 模型可以自动提取并学习到比传统机器学习模型更多的分类特征,这也说明了深度学习模型在特征自动提取和学习方面相比于传统机器学习更有优势.而 CNN 模型分类效果在引入注意力机制后有了一定的提升,原因在于增加了注意力机制提取的局部特征,说明注意力机制对文本分类效果提升起到了一定的作用.

从表 2 可以看出,随着窗口大小的增加,利用 Attention-CNN 分类模型分类效果会随之增长.当窗口大小为 1 时,取得的分类效果最差,因为当窗口大小为 1 时,没有考虑到上下文的词对中心词的影响;当窗口大小为 5 时,分类效果最好.

5　结语

本文利用 word2vec 训练大规模中文新闻语料,从大量的文本信息中得到中文词的词向量,作为文本的特征表达.相比于传统的机器学习提取特征的方法,word2vec 可以自动将语义信息浓缩进数学向量中,以向量作为分类模型的输入特征.本文在经典的 CNN 分类模型的基础上,引入了注意力机制,在全局特征上又加入了局部特征,这样使特征表达更加丰富.从实验结果也可以看出,引入注意力机制后的 CNN 模型对文本的分类效果有了一定程度的提升.

参 考 文 献

[1] Zhang Y, Jin R, Zhou Z H. Understanding bag-of-words model: a statistical framework [J]. International Journal of Machine Learning and Cybernetics, 2010, 1(1-4): 43-52.

[2] Collobert R, Weston J, Bottou L, et al. Natural Language Processing (Almost) from Scratch [J]. Journal of Machine Learning Research, 2011, 12(1):2493-2537.

[3] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [EB/OL]. (2013-09-07) [2017-05-08]. https://arxiv.org/abs/1301.3781.

[4] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [C]//MIT Press. Advances in Neural Information Processing Systems. Massachusetts: MIT Press, 2013: 3111-3119.

[5] Kim Y. Convolutional neural networks for sentence classification [EB/OL]. (2014-09-14) [2017-04-13]. https://arxiv.org/abs/1408.5882.

[6] Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification [EB/OL]. (2016-04-06) [2017-05-10]. https://arxiv.org/abs/1510.03820.

[7] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507.

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

[9] Hubel D H, Wiesel T N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex [J]. Journal of Physiology, 1962, 60(1): 106-154.

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

[11] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks [C]// Springer International Publishing. European Conference on Computer Vision. Berlin: Springer International Publishing, 2014: 818-833.

[12] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-04-10) [2017-05-12]. https://arxiv. org/abs/1409.1556.

[13] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions [C]// IEEE. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE. 2015: 1-9.

[14] Mnih V, Heess N, Graves A. Recurrent models of visual attention [C]//MIT Press. Advances in Neural Information Processing Systems. Massachusetts: MIT Press, 2014: 2204-2212.

[15] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [EB/OL]. (2016-05-19) [2017-05-04]. https://arxiv.org/abs/1409.0473.

[16] Luong M T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation [EB/OL]. (2015-09-20) [2017-05-07]. https://arxiv.org/abs/1508.04025.

[17] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research, 2014, 15(1):1929-1958.

[18] Tong S, Koller D. Support vector machine active learning with applications to text classification [J]. Journal of Machine Learning Research, 2001, 2(11): 45-66.

[19] Han E H S, Karypis G, Kumar V. Text categorization using weight adjusted *k*-nearest neighbor classification [C]// Springer International Publishing. Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin: Springer International Publishing, 2001: 53-65.

[20] McCallum A, Nigam K. A comparison of event models for naive Bayes text classification [C]//AAAI. AAAI-98 Workshop on Learning for Text Categorization. New York:AAAI,1998, 752: 41-48.

(责任编辑　曹　东)