计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(7):145−150 [doi: 10.15888/j.cnki.csa.006972]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

针对文本分类的神经网络模型①

涂文博, 袁贞明, 俞　凯

(杭州师范大学 信息科学与工程学院, 杭州 311121)
(移动健康管理系统教育部工程研究中心, 杭州 311121)
通讯作者: 袁贞明, E-mail: zmyuan@hznu.edu.cn

摘　要: 文本分类是自然语言处理领域的一项重要任务, 具有广泛的应用场景, 比如知识问答、文本主题分类、文本情感分析等. 解决文本分类任务的方法有很多, 如支持向量机 (Support Vector Machines, SVM) 模型和朴素贝叶斯 (Naïve Bayes) 模型, 现在被广泛使用的是以循环神经网络 (Recurrent Neural Network, RNN) 和文本卷积网络 (TextConventional Neural Network, TextCNN) 为代表的神经网络模型. 本文分析了文本分类领域中的序列模型和卷积模型, 并提出一种组合序列模型和卷积模型的混合模型. 在公开数据集上对不同模型进行性能上的对比, 验证了组合模型的性能要优于单独的模型.

关键词: 文本分类; 自然语言处理; 神经网络

引用格式: 涂文博,袁贞明,俞凯.针对文本分类的神经网络模型.计算机系统应用,2019,28(7):145–150. http://www.c-s-a.org.cn/1003-3254/6972.html

Neural Network Models for Text Classification

TU Wen-Bo, YUAN Zhen-Ming, YU Kai

(School of Information Science and Engineering, Hangzhou Normal University, Hangzhou 311121, China)
(Engineering Research Center of Mobile Health Management System, Ministry of Education, Hangzhou 311121, China)

Abstract: Text classification is an important task in the field of natural language processing. It has a wide range of applications, such as knowledge question and answer, text topic classification, text emotion analysis, and so on. There are many methods to solve the task of text classification, such as Support Vector Machines (SVM) model and Naïve Bayes model. Typical neural network models widely used now are the Recurrent Neural Network (RNN) and the Text Conventional Neural Network (TextCNN). In this study, the sequence model and convolution model in the field of text classification are analyzed, and a hybrid model of combining sequence model and convolution model is proposed. By comparing the performance of the different models on the open dataset, it is proved that the performance of the combined model is better than that of the single model.

Key words: text-classification; natural language processing; neural network

在自然语言处理 (Natural Language Processing, NLP) 领域, 文本分类是一项重要任务, 具有广泛的应用场景, 比如知识问答、文本主题分类、文本情感分析等. 很多专家学者提出不同方法来解决文本分类问题, 文献[1]提出基于规则特征的支持向量机 (Support Vector Machines, SVM) 模型对问答系统的问题进行分类, 文献[2]朴素贝叶斯 (Naïve Bayes) 与 SVM 相结合, 提出了一种简洁高效的情感和主题分类模型. 文献

① 基金项目: 浙江省自然科学基金 (LQ16H180004); 杭州市科技计划项目 (20162013A02)
Foundation item: Natural Science Foundation of Zhejiang Province (LQ16H180004); Science and Technology Plan of Hangzhou Municipality (20162013A02)
收稿时间: 2019-01-05; 修改时间: 2019-01-24; 采用时间: 2019-01-31; csa 在线出版时间: 2019-06-28

[3]使用三支决策方法提取文本特征, 提高了基于规则特征的文本分类模型的准确率. 随着深度学习技术发展迅速, 不同的神经网络模型开始被应用到文本分类任务当中, 以循环神经网络 (Recurrent Neural Network, RNN)[4]为代表的序列模型 (Sequence Models) 被大规模使用[5,6]. 文献[7]把注意力模型添加到序列模型中, 完成文本分类任务并获得比较好的结果. 文献[8]将一直用于图像处理的卷积神经网络 (Conventional Neural Network, CNN)[9]首次应用到文本分类领域, 并获得较好的结果. 之后, 文献[10–12]将卷进神经网络模型做了不同程度的改变, 用于文本分类任务. 本文对文本分类领域使用的 RNN 模型和 TextCNN 模型分别作介绍, 并提出一种组合 RNN 和 TextCNN 的混合模型, 在公开数据集上对这几种模型进行对比, 以评估不同类型模型的性能, 并验证了组合模型的性能要优于单独的模型.

1 文本表示

文本表示, 即将字符文本通过某种形式表示成计算机可以处理的数值化数据. 由于机器学习和深度学习的算法模型都不能直接处理字符文本, 所以在使用机器学习或深度学习模型做 NLP 任务时, 需要将原始的字符文本做数值化表示, 转换成数值向量. 并且, 不同的文本表示形式对算法模型结果的影响也有所不同. 根据 NLP 任务和算法模型的不同, 可以将文本的字或词作为最小表示单位 (本文中, 以字作为文本的最小表示单位), 进行数值向量化.

文本表示方法大致分为两种, 一种是 One-Hot 式编码. 这种方式是从待数值化的所有文本中建立一个全局的完备字典, 该字典包含文中出现的所有字, 用字典序的数字来表示一个字. 采用这种方法编码, 字典大小就是字的向量维度. 在向量中, 只有该字的字典序对应位置上的数字为 1, 其他位置均为 0. 这种方法有两个主要缺点, 一是字的向量十分稀疏, 二是字与字之间没有任何关联, 比如它并不能表现出同性字或意义相近的之间的相似性. 另外一种方法是分布式表示法 (Distributed Representation)[13], 它基于“上下文相似的词, 其语义也相似”这一假说, 其基本思想是使用统计学方法通过训练把句子中的每个字映射成 K 维的实数向量, 通过字与字的实数向量之间的距离 (如欧氏距离、余弦距离等) 来描述字之间的语义相似性, 即相似语义的字具有相似的数值向量. 现在的 NLP 任务中, 使用的大多是基于分布式表示法的文本表示模型. 代表的有文献[14]提出的 Word2Vc 模型和斯坦福大学提出的 GloVe 模型[15]. 在本文的实验部分, 采用 Word2Vec 模型训练的字向量作为文本分类模型的输入数据.

2 文本分类神经网络模型

2.1 LSTM

循环神经网络 (Recurrent Neural Network, RNN) 是 NLP 任务中被广泛使用的网络模型. 它通过循环使用同一个网络神经元来处理任意长度的序列, 将序列前面的特征传递给后面作为输入, 从而捕获到完整的序列上下文特征信息用于序列分类或序列标注.

当序列长度较长时, 神经网络参数的训练过程会遇到梯度爆炸和梯度消失问题[15–17], 造成模型的参数学习缓慢甚至停止学习. 文献[17]提出 LSTM 模型, 对 RNN 做了改进, 其目的是解决 RNN 存在的梯度爆炸问题. LSTM 的神经元结构如图 1 所示.

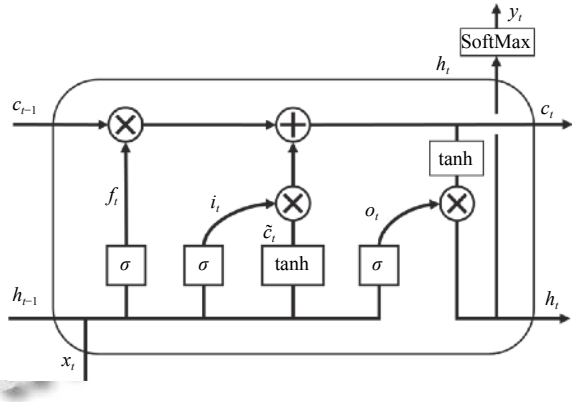

图 1 LSTM 神经元结构

定义, 序列的 t 位置为时刻 t, LSTM 神经元由三个“门”控制: 遗忘门 f_t, 输入门 i_t 和输出门 o_t. 三个门的值为 0 或 1. 在时刻 t, LSTM 的参数更新方式如下:

$$\widetilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{1}$$

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{2}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3}$$

$$o_t = \sigma(W_o [h_{t-1}, x_t] + b_o) \tag{4}$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{5}$$

$$h_t = o_t * \tanh(C_t) \tag{6}$$

其中, x_t 是 t 时刻的输入向量, h_{t-1} 为 t-1 时刻的输出向

量. 输入门 i_t 控制是否将当前时刻信息存放在记忆细胞 C 中; 遗忘门 f_t 控制是否使用记忆细胞的历史信息; 输出门决定是否让当前记忆细胞的信息参与当前时刻输出值的计算. W_C、W_f、W_i、W_o 分别为隐藏单元的权重矩阵, b_C、b_f、b_i、b_o 分别为偏置矩阵. σ 和 tanh 分别为 Sigmoid 函数和 tanh 函数, 定义由公式 (7) 和公式 (8) 给出.

$$f(x)=\frac{1}{1+e^{-x}} \tag{7}$$

$$\tanh(z)=\frac{e^{z}-e^{-z}}{e^{z}+e^{-z}} \tag{8}$$

BiLSTM 是使用两个 LSTM 网络分别从前向和后向处理序列, 这样能更全面地提取序列的上下文信息.

本文的实验部分, 采用 BiLSTM 作为 RNN 的替代模型, 用来与其他模型对比. 使用 BiLSTM 处理文本分类的基本流程结构如图 2 所示.

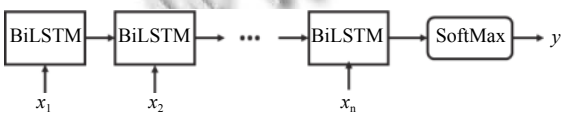

图 2 LSTM 模型用于文本分类示意图

其中, x_i 为字向量, y 为文本序列的类别. SoftMax 是用于多分类的激活函数, 假设 $Z=[z_1, z_2, \cdots, z_n]$是一个 n 维的特征向量, 其 SoftMax 值由下式给出:

$$s(Z)_i=\frac{e^{z_i}}{\sum_{k=1}^{K}e^{z_k}} \tag{9}$$

2.2 TextCNN

卷积神经网络 (Conventional Neural Network, CNN) 被广泛应用于计算机视觉中. 随着词嵌入和深度学习技术的发展,现在很多学者开始在 NLP 任务中使用 CNN.

卷积操作是 CNN 的重要特征之一. 卷积层以特征映射为组织方式, 其中的每一个单位与前一层的局部感受野相连接, 利用共享的卷积核 (或称过滤器) 与局部感受野做卷积运算, 再经过激活函数 (如 $ReLUv$、$tanh$) 做非线性运算, 得到特征值. 给定一个矩阵 $\boldsymbol{X}\in\mathbb{R}^{M\times N}$, 和卷积核 $\boldsymbol{F}\in\mathbb{R}^{M\times N}$, 一般 $m\ll M, n\ll N$, 其卷积如式 (10) 所示:

$$conv_{ij}=\sum_{u=1}^{m}\sum_{v=1}^{n}f_{uv}\cdot x_{i-u+1:j-v+1}. \tag{10}$$

图 3 是文献[8]提出用于文本分类的 TextCNN 模型结构示意图, 在本文的实验部分, 使用该模型作为 CNN 在文本分类领域的代表模型与其他模型对比.

TextCNN 通过不同的通道数目和卷积核大小, 使用一维卷积的方式提取句子矩阵的特征. Max-over-time 池化层的作用是从提取的特征矩阵中选出最大值, 跟其他通道的最大值拼接, 组合成筛选过的特征向量, 继而通过 SoftMax 层对文本进行分类.

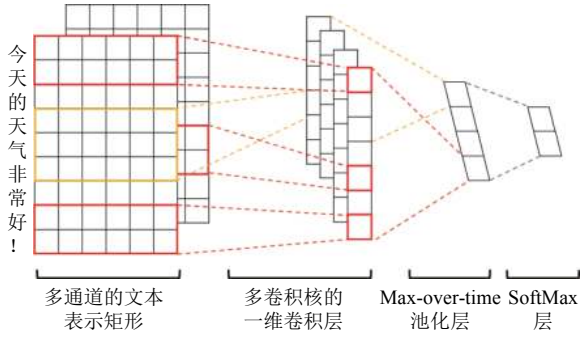

图 3 TextCNN 模型示意图

3 BiLSTM+CNN 模型

鉴于以上描述, BiLSTM 和 CNN 在应对文本分类问题时, 各有特色, BiLSTM 由于具有自动学习记忆文本序列特征的特点, 对于文本特征的提取、语音的理解和长文本依赖问题有很好的适用性; 而 TextCNN 凭借不同的卷积核和通道数目, 比较适合提取更复杂的文本特征. 我们尝试将这两个模型进行组合, 组成的混合模型结构如图 4 所示.

在 BiLSTM 模型中, 其每一个神经元都可以有输出, 该输出表示为句子截至到该字符时, 网络模型提取到的句子特征. 该法提取的句子特征为二维张量, 适合 TextCNN 做深层次的特征提取.

BiLSTM+TextCNN 组合模型的思想是, 将 BiLSTM 的每一个神经元输出的特征连结成句子的特征矩阵, 用作卷积神经网络的输入, 进行特征的二次提取. 具体的, 句子中的字组成的字向量序列, 经过 BiLSTM 模型编码成二维矩阵, 编码的过程即递归模型学习的过程, 该过程将句子的字向量特征过滤、融合成句子的特征矩阵. 递归模型的特点在于捕获长距离的文本语义特征, 对长距离的语义进行关联特征提取, 而 BiLSTM 模型保证了句子的前后向语义信息都可以被捕获. 经过 BiLSTM 模型输出的句子特征矩阵包含了句子中字与字之间长距离的语义信息, 使用 TextCNN 对句子的特征矩阵做卷积操作, 并通过池化进行特征筛选, 继而提取句子特征矩阵中相邻文字的

关联特征, 这样相互结合, 使得句子的语义特征得以全面且深度的提取, 从而可以获得更准确的分类结果. BiLSTM+TextCNN 组合模型的参数学习过程如下:

算法 1. BiLSTM+TextCNN 模型参数学习过程

1) 随机初始化模型的参数, 设置模型批处理数据量 batchsize 大小和迭代次数 epochs.
2) 将字向量表示的句子输入到 BiLSTM 网络中, 获得经 BiLSTM 提取的句子特征矩阵.
3) 将特征矩阵输入到 TextCNN 模型中, 进行文本特征的二次提取.
4) 将第 3) 步由 TextCNN 提取的特征输入到全连接网络层和 SoftMax 层中获得文本的分类结果.
5) 经由代价函数计算模型中参数的梯度, 进行反向传播, 更新参数值.
6) 重复 2)–5) 步直至满足设定的 epochs 次数止.

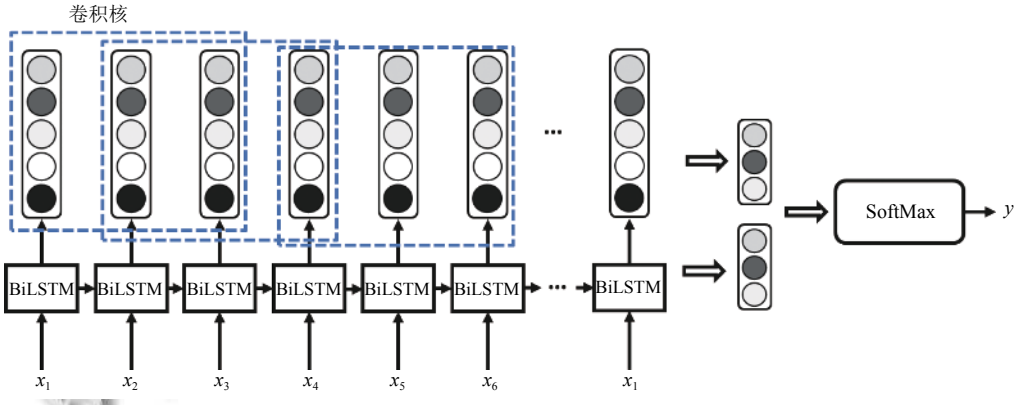

图 4 BiLSTM+CNN 模型示意图

模型的代价函数是模型参数学习过程中的目标函数或准则. 通过最小化代价函数来优化模型, 获得更准确的拟合参数. 模型采用交叉熵损失函数, 其表示如式 (11) 所示.

$$C = -\frac{1}{n}\sum_{i=1}^{n} y\log a + (1-y)\log(1-a) \tag{11}$$

其中, n 为分类任务中的类别数目, y 为数据集中的真实标签值, a 为模型经过学习后预测的标签值. 模型参数学习的过程就是在最小化该损失函数.

4 实验

4.1 数据集

实验采用的数据集是由清华大学自然语言处理与社会人文计算实验室公开的新闻文本数据集[1], 该数据集包含 836 075 条新闻文本, 共 14 个候选分类类别: 体育、财经、房产、家居、彩票、教育、科技、股票、时尚、时政、星座、游戏、社会、娱乐. 由于星座和社会的数据量过少, 在实验中略去该两类数据, 使用原数据集中的 12 个类别的新闻文本进行实验, 共计 824 900 条. 数据集划分百分之八十为训练集, 百分之二十为测试集.

4.2 实验设置

实验设置为使用上述数据集进行文本分类, 并评估效果. 作为对比, 实验增加一个 BP 神经网络 (Back Propagation Neural Network, BPNN) 模型作为分类任务的基准, 用以对比 BiLSTM 模型、使用 TextCNN 模型和本文提出的 BiLSTM+TextCNN 的组合模型的性能. 模型的主要超参数如表 1 所示.

4.3 评测方法

模型采用 $F1$ 值作为评估标准. $F1$ 值由查准率 (Precision, P) 和查全率 (Recall, R) 经过计算获得. 定义 $\boldsymbol{y}$ 为模型输出的字标签预测分类值集合, $\hat{y}$ 为数据集字标签的真实值集合, 模型的查准率 $P(y,\hat{y})$ 由式 (12) 给出定义:

$$P(y,\hat{y}) = \frac{|y \cap \hat{y}|}{|y|} \tag{12}$$

查全率 $R(y,\hat{y})$ 由下式给出:

$$R(y,\hat{y}) = \frac{|y \cap \hat{y}|}{|\hat{y}|} \tag{13}$$

$F1$ 值由式 (14) 给出定义, 它是 P 值和 R 值的调和

[1] 数据集版权归清华大学自然语言处理与社会人文计算实验室所有. 详见 http: //thuctc.thunlp.org/

平均数, 其中 β=1.

$$F_\beta(y,\hat{y})=\left(1+\beta^2\right)\frac{P(y,\hat{y})\times R(y,\hat{y})}{\beta^2\times P(y,\hat{y})+R(y,\hat{y})} \quad (14)$$

$F1$ 值越高, 模型表现越好.

表 1 实验中模型采用的主要超参数

超参数	值
字向量维度	100
BPNN 隐藏层单元数	200
BPNN 隐藏层数	2
BiLSTM 层数	2
BiLSTM Dropout	0.7
BiLSTM 学习率	0.001
BiLSTM Batch Size	128
TextCNN 卷积核大小	5
TextCNN 通道数	100
TextCNN 学习率	0.001
TextCNNBatchSize	256
组合模型 BiLSTM 层数	2
组合模型 Dropout	0.5
组合模型 BatchSize	128
Epoch	20

4.4 实验结果

经过多次实验, 几个模型在不同类别的分类表现如表 2 所示. 表中列举了几个模型对所有 12 个分类的 $F1$ 值结果 (数值为多次实验数据中, 最高三次的均值). 从结果中可以看出, BPNN 模型由于结构较为简单, 其性能整体落后于其他三个模型, BiLSTM 模型在体育、财经、房产、股票、社会这五个分类上与 TextCNN 模型的表现较为接近, 但在其他分类上要弱于 TextCNN 模型. 而本文提出的组合 BiLSTM 和 TextCNN 的模型, 其表现整体上要好于其他两个单独的模型.

究其原因, BPNN 模型结构较为简单, 只是两层全连接网络的组合, 对文本的上下文信息的特征捕获能力不强, 做不到字与字之间关联语义特征的提取, 因而整体性能要弱于业界较为通用的文本分类模型. 不过由于其模型结构简单, 在运行速度上要优于其它模型. BiLSTM 模型由于循环使用神经元学习序列中潜在的关联特征, 并选择性记忆序列中的语义信息, 对文本的上下文特征捕获能力较强, 自然获得较高的分类准确率. TextCNN 模型使用不同的卷积核, 对文字序列的相邻语义特征捕获能力较强, 并能通过多层卷积融合较长距离的语义关系, 因而可以得到较好的表现. 并且, CNN 模型很适合并行计算, 借由 GPU 的加速, TextCNN 模型的训练时间要比 BiLSTM 快很多. 本文提出的 BiLSTM+TextCNN 模型, 结合了 BiLSTM 模型和 TextCNN 模型的特点, 使用 BiLSTM 对文本序列的上下文特征做筛选和提取, 再经 TextCNN 进行更进一步、更细粒度的特征选择与融合, 将句子中的长、短距离文字的关联信息和句子的语义信息较为全面的捕获, 故其性能表现要优于单独的 BiLSTM 模型或 TextCNN 模型. 不过由于组合了两个复杂模型, 混合模型在时间复杂度上要差一些.

另外, 除了 BPNN 模型外, 其他模型的 $F1$ 值均超过了 0.9, 并有过半数目的分类问题 $F1$ 值超过了 0.95, 这说明几个模型在文本分类问题上都能够得到较为准确的结果, 并且将序列模型和卷积模型组合会得到更佳的效果.

表 2 不同神经网络模型在数据集上的分类效果 $F1$ 值对比

模型	类别											
	体育	财经	房产	家居	教育	科技	股票	时尚	时政	游戏	社会	娱乐
BPNN	0.925	0.931	0.929	0.877	0.892	0.904	0.901	0.921	0.889	0.930	0.912	0.932
BiLSTM	0.972	0.975	0.986	0.919	0.920	0.943	0.929	0.951	0.927	0.955	0.936	0.962
TextCNN	0.976	0.974	0.986	0.931	0.946	0.968	0.926	0.973	0.952	0.981	0.939	0.977
BiLSTM+TextCNN	0.984	0.981	0.987	0.939	0.957	0.973	0.930	0.969	0.966	0.989	0.948	0.984

5 总结和展望

文本分类问题一直是 NLP 领域关注的重点, 有很广泛的应用场景. 在处理文本分类任务时, 主流的模型有以 BiLSTM 为代表的序列模型和以 TextCNN 为代表的卷积模型. 本文介绍了这两个主流的文本分类模型, 并提出新的组合模型, 通过实验对比了不同模型的性能表现. 未来, 将多种模型组合起来解决某一类问题或是后续研究的重点.

参考文献

1 Silva J, Coheur L, Mendes AC, *et al*. From symbolic to sub-symbolic information in question classification. Artificial Intelligence Review, 2011, 35(2): 137–154. [doi: 10.1007/s10462-010-9188-4]

2 Wang SD, Manning CD. Baselines and bigrams: Simple, good sentiment and topic classification. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers. Jeju Island, Korea. 2012. 90–94.
3 靳义林, 胡峰. 基于三支决策的中文文本分类算法研究. 南京大学学报 (自然科学), 2018, 54(4): 794–803.
4 Funahashi KI, Nakamura Y. Approximation of dynamical systems by continuous time recurrent neural networks. Neural Networks, 1993, 6(6): 801–806. [doi: 10.1016/S0893-6080(05)80125-X]
5 Socher R, Huval B, Manning CD, *et al*. Semantic compositionality through recursive matrix-vector spaces. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, Korea. 2012. 1201–1211.
6 Liu PF, Qiu XP, Huang XJ. Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv: 1605. 05101, 2016.
7 郑雄风, 丁立新, 万润泽. 基于用户和产品 Attention 机制的层次 BGRU 模型. 计算机工程与应用, 2018, 54(11): 145–152. [doi: 10.3778/j.issn.1002-8331.1701-0337]
8 Kim Y. Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1746–1751.
9 Sahiner B, Chan HP, Petrick N, *et al*. Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images. IEEE Transactions on Medical Imaging, 1996, 15(5): 598–610. [doi: 10.1109/42.538937]
10 Wang ZH, Wang XX, Wang G. Learning fine-grained features via a CNN tree for large-scale classification. Neurocomputing, 2018, 275: 1231–1240. [doi: 10.1016/j.neucom.2017.09.061]
11 Yang X, Macdonald C, Ounis I. Using word embeddings in twitter election classification. Information Retrieval Journal, 2018, 21(2-3): 183–207. [doi: 10.1007/s10791-017-9319-5]
12 陈珂, 梁斌, 柯文德, 等. 基于多通道卷积神经网络的中文微博情感分析. 计算机研究与发展, 2018, 55(5): 945–957. [doi: 10.7544/issn1000-1239.2018.20170049]
13 Hinton GE. Learning distributed representations of concepts. Proceedings of the Eighth Annual Conference of the Cognitive Science Society. Amherst, Massachusetts, UK. 1986. 1–12
14 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2013. 3111-3119.
15 Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar. 2014. 1532-1543.
16 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]
17 Hochreiter S, Bengio Y, Frasconi P, *et al*. Gradient flow in recurrent nets: The difficulty of learning long-term dependencies. Kolen JF, Kremer SC. A Field Guide to Dynamical Recurrent Networks. Los Alamitos: IEEE Press, 2001.