

基于深度学习的文本分类研究进展

杜思佳, 于海宁, 张宏莉

(哈尔滨工业大学计算机科学与技术学院, 黑龙江 哈尔滨 150001)

摘要: 文本分类技术是自然语言处理领域的研究热点, 其主要应用于舆情检测、新闻文本分类等领域。近年来, 人工智能技术在自然语言处理的许多任务中有着很好的表现, 将神经网络技术应用于文本分类取得了许多成果。在基于深度学习的文本分类领域, 文本分类的数值化表示技术和基于深度学习的文本分类技术是两个重要的研究方向。对目前文本表示的有关词向量的重要技术和应用于文本分类的深度学习方法的实现原理和研究现状进行了系统的分析和总结, 并针对当前的技术发展, 分析了文本分类方法的不足和发展趋势。

关键词: 文本分类; 深度学习; 人工神经网络; 词向量

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2020010

Survey of text classification methods based on deep learning

DU Sijia, YU Haining, ZHANG Hongli

School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

Abstract: Text classification is a research hot spot in the field of natural language processing, which is mainly used in public opinion detection, news classification and other fields. In recent years, artificial neural networks has good performance in many tasks of natural language processing, the application of neural network technology to text classification has also made many achievements. In the field of text classification based on deep learning, numerical representation of text and deep-learning-based text classification are two main research directions. The important technology of word embedding in text representation and the implementation principle and research status of deep learning method applied in text classification were systematically analyzed and summarized. And the shortcomings and the development trend of text classification methods in view of the current technology development were analyzed.

Key words: text classification, deep learning, artificial neural network, word embedding

收稿日期: 2019-11-25; 修回日期: 2020-01-25

通信作者: 于海宁, yuhaining@hit.edu.cn

基金项目: 国家自然科学基金 (61601146, 61732022)

Foundation Item: The National Natural Science Foundation of China (61601146, 61732022)

论文引用格式: 杜思佳, 于海宁, 张宏莉. 基于深度学习的文本分类研究进展[J]. 网络与信息安全学报, 2020, 6(4): 1-13.
DU S J, YU H N, ZHANG H L. Survey of text classification methods based on deep learning[J]. Chinese Journal of Network and Information Security, 2020, 6(4): 1-13.

1 引言

文本分类是对文本集按照一定的分类体系或标准进行自动分类标注的技术,主要应用于信息检索、垃圾文本过滤、舆情检测、情感分析等领域,一直是自然语言处理领域的研究热点^[1-2]。文本分类方法经历了基于词匹配、基于知识工程、基于统计和机器学习的发展历程。近年来,深度学习在文本分类等自然语言处理任务中的研究和应用得到了学术界的广泛关注,并且取得了一些重大的进展。以非结构化的字符数据为内容的自然语言文本,在进行下游的分类任务之前,要将其转化为数值化表示的可计算的数据。因此,作为词的分布式表示方法,词向量的提出和发展对于文本分类任务的研究具有重要的意义。

目前,在基于深度学习的文本分类方法方面,有两种侧重的研究方向:一种专注于词向量模型的研究,致力于训练高质量的词向量;另一种侧重于文本分类方法即深度神经网络分类模型的研究,其中比较有代表性的神经网络结构为卷积神经网络(CNN, convolutional neural network)和循环神经网络(RNN, recurrent neural network)。本文从基于深度学习的词向量训练模型和文本分类算法的研究和发展,这两个基于深度学习的文本分类的研究重点来介绍现在主流的文本分类方法,并对现有研究的不足进行分析。

2 词向量模型的研究现状

词向量是一种词语的稠密向量化表示方法,其早期概念 distributed representation 由 Hinton 等^[3]在 1986 年提出。在词向量的分布式表示概念出现之前,独热编码(one-hot encoding)是在自然语言处理中被广泛采用的形式,one-hot 编码将词语表示为词表大小维度的长向量,其中只有一个维度的值为 1,代表当前词,其他维度均为 0。one-hot 编码采用稀疏方式表示词向量,表示简单、易于理解,但存在很多问题。首先,在这种编码下,任意两个词是孤立的,它们之间的相关性无法体现;其次,由于词表的数量级,one-hot 编码的词向量维度非常大,在解决某些任务时,可能会造成维数灾难^[4]。词向量以稠密向量的方式表示词

语,首先,使向量维度相对 one-hot 编码有了显著降低,在一定程度上缓解了维数灾难的问题;其次,实践证明,经过训练得到的较为优秀的词向量,可以通过计算向量之间欧几里得距离的简单方式来衡量词语之间的相似性,这说明,通过训练的以稠密向量方式表示的词向量在一定程度上代表了词语所具有的潜在含义。

语言模型是根据语言客观事实进行的语言抽象数学建模,广泛地应用在自然语言处理的各项任务中。简单来说,语言模型就是给定一个句子,计算其是自然语言的概率 $P(w_1, w_2, \dots, w_t)$, 其中, w_1, w_2, \dots, w_t 依次表示句子中的各个词。在实际求解中,常用的推论是 $P(w_1, w_2, \dots, w_t) = P(w_1)P(w_2 | w_1) \dots P(w_t | w_1, \dots, w_{t-1})$ 。由于上述推论中参数过多、计算繁杂,常见的各种语言模型在近似求 $P(w_t | w_1, \dots, w_{t-1})$ 的定义上有所不同。而词向量则是在训练语言模型时得到的产物。

现在主流的基于神经网络的词向量训练方法以大量无标注文本数据进行无监督学习得到词向量。以神经网络训练词向量模型影响最为深远的是 Bengio 等^[4]提出的神经网络语言模型(NNLM, neural network language model),其网络结构如图 1 所示。NNLM 基于 n -gram 语言模型,使用 3 层神经网络,以前 $n-1$ 个词作为输入预测第 n 个词。其神经网络为线性网络,隐藏层采用 tanh 为激活函数,输出层以 softmax 为激活函数。这种词向量训练方法奠定了词向量训练的技术方向,之后的研究人员在词向量训练方法上的研究多以此为启发。Collobert 等^[5]提出将词向量用于解决自然语言处理中词性标注、命名实体识别等各项任务的思想,并对此做了多项系统工作,发表了其开源系统 SENNA。同时,在词向量训练方面,以中心词及其上下文作为输入,输出中心词与上下文关联程度的得分。Mnih 等^[6]提出了 log 双线性语言模型(LBL, log-bilinear language model),后来在此基础上又将层级的思想引入词向量训练,提出层级 softmax 函数,在保证效果的前提下大大提升了训练速度^[7-8],在 2013 年采用向量点积结构提出了基于向量的逆语言模型(ivLBL)^[9]。Mikolov 等^[10-11]将循环神经网络引入词向量训练,循环神经网络由于其时序特点,可以充分利用上下文信息来预测下一个词,但其

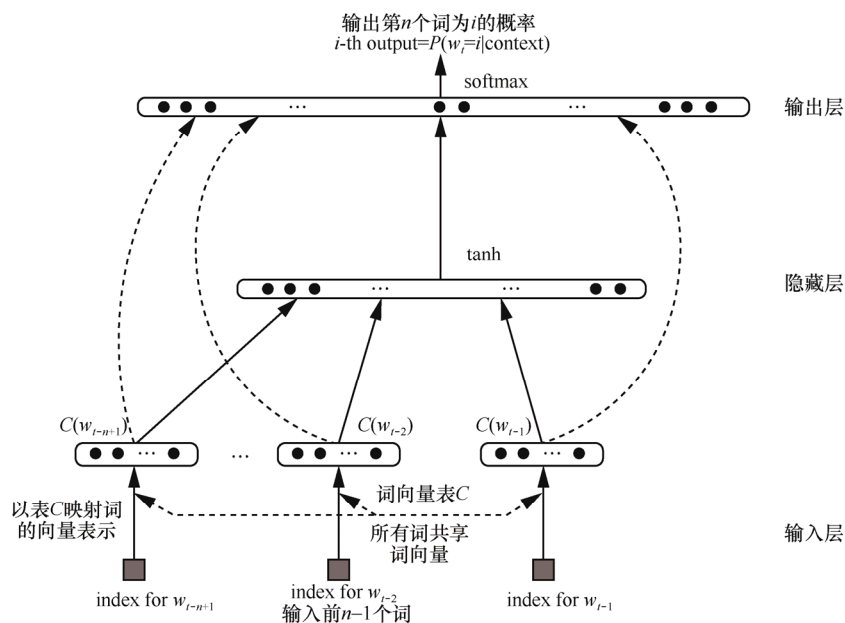


图1 NNLM 网络结构
Figure 1 NNLM network structure

主要思想依旧依托于 n -gram 语言模型。针对词向量的训练, 研究人员还进行了更多的探索。但是, 其主要思想大同小异, 大多从 Bengio 等提出的 NNLM 出发, 以上下文预测为主要思想, 在神经网络模型的选择和具体实现、语言模型的训练速度等方面进行创新和改进。

自然语言处理领域最为常用的词向量训练方法是 2013 年 Google Mikolov 提出的 Word2vec 及其发布的相应工具包^[10]。Word2vec 有两种可供选择的语言模型 (CBOW 和 Skip-gram), 并且提供了两种有效提高训练速度的优化方法: 层次 softmax (hierarchical softmax) 和负采样 (negative sampling) 技术。Word2vec 在之前研究者的基础上, 简化模型, 取其精华, 得到现有模型, 其在百万数量级的词表和语料中均能进行高效的训练, 且获取的词向量能够较为精确地衡量词与词之间的相似性。CBOW 的主要思想是以一个词的上下文作为输入, 预测当前词出现的概率; 而 Skip-gram 相反, 它以这个词作为输入, 预测其上下文。两种方法均可自定义上下文的窗口大小。层次 softmax 是以哈夫曼树代替神经网络从隐藏层到 softmax 层的映射, 这样可以有效提高模型训练速度。负采样技术在词向量训练中引入了负例, 而其优化目标为最大化正样本的似

然, 同时最小化负样本的似然。另一个常用的词嵌入工具是由斯坦福 NLP 团队发布的 Glove^[12], 它是一个基于全局词频统计的词表征工具。Glove 模型融合了传统机器学习的全局词-文本矩阵分解方法和以 Word2vec 为代表的局部文本框捕捉方法, 在有效利用词的全局统计信息的同时, 有效捕捉其上下文信息。Facebook 发布的工具——Fast Text 中也有词向量训练的部分, 其基本模型与 Word2vec 一致, 不同之处在于, FastText 的词向量训练在输入层位置提供了字符级别的 n -gram, 如单词为 where, $n=3$ 时, 其 n -gram 为 <wh,whe,her,ere,re>, 单词将由单词本身及其 n -gram 代表。由于英文单词具有前后缀等语言形态上的相似性, 这种方法可以建立词与词之间的联系^[13]。

前面所介绍的词向量模型, 包括被广泛使用的 Word2vec、Glove 等, 其生成的词向量均存在一个问题: 在利用这些词向量对词语进行编码的时候, 一个词只对应一个词向量, 即无法解决词的多义性问题。为解决这个问题, ELMo 提供了一种解决方案^[14], 实验结果表明, ELMo 在分类任务、阅读理解等 6 个 NLP 任务中性能有不同幅度的提升, 最高的提升达到 24.9%。ELMo 是一种新型深度语境化词表征, 其主要思想是: 事先训练好一个完整的语言模型, 然后在实际使用时

根据特定上下文单词的语义调整词向量。因此 ELMo 模型对同一个词在不同的上下文环境中能够生成不同的词向量。预训练过程中, ELMo 的语言模型任务是以词的上下文预测这个词, 其网络结构采用双向 LSTM, 其中, 正向 LSTM 编码器输入为正序的词的上文, 逆向 LSTM 编码器输入为逆序的词的下文。在 Fine-Tuning 阶段, 一般先用下游任务的语料对预训练的模型进行再训练, 对词向量进行调整, 再将其应用于分类任务。对于输入句子中的每个单词, 均能得到叠加的 3 个向量, 即第一层单词的词向量和双向 LSTM 中单词对应位置的向量, 其中每一个词向量都学到了不同的语义信息。在下游任务中, 以一定的权重将 3 个词向量累加求和得到对应单词的词向量, 其中权重可通过学习得到, 然后将此词向量作为下游分类任务的网络结构中对应单词的词向量输入。OpenAI 在 2018 年提出的 GPT 模型^[15]是进行语言模型预训练, 再在 Fine-Tuning 模式解决下游任务的两阶段模型。其与 ELMo 模型不同之处在于, 首先, GPT 模型的下游任务不再是自己设计的网络结构, 而要将其改造成和 GPT 一样的网络结构, 即 Fine-Tuning 的思想; 其次, 在特征抽取器的选择上, GPT 采用了 Google 团队 2017 年提出的 Transformer 模型^[16], Transformer 模型是采用叠加的自注意力 (self-Attention) 机制构成的深度网络, 拥有很强的特征提取能力。此外, 不同于 ELMo 以上下文预测单词的思想, GPT 采用单向的语言模型, 仅以词的上文来进行预测。同年, Google 发布的语言模型预训练方法 Bert^[17]在 11 个各种类型的 NLP 任务中都达到了目前最好的效果。Bert 模型采用和 GPT 相同的两阶段模型, 主要不同之处在于 Bert 采用了双向语言模型, 并且语言模型的数据规模比 GPT 更大。另外, 在训练方法上, 提出了两个新的目标任务: 遮挡语言模型 (MLM, masked language model) 和预测下一个句子 (next sentence prediction) 的任务。

在应对多义词问题上, Facebook 在其发布的开源词向量及文本分类工具 FastText 中采用的词级别的 n -gram 方法值得关注^[18]。由于其分类方法采用句子的词向量均值作为输入, 以简单线性分

类器作为分类器, 为了在输入中保留句子的上下文序列信息, FastText 利用了词级别的 n -gram 进行词向量表征, 这与前文提到的 FastText 词向量训练中字符级别的 n -gram 有所不同。具体来说, 若 $n=2$, 词序列为 a,b,c, 则其 n -gram 为 <ab,abc,bc>, 之后将所有的 n -gram 和原词的词向量相加来代表上下文信息, 由于 n -gram 中含有上下文信息, 在一定程度上解决了多义词问题。但这种解决方案相对于 ELMo、GPT 和 Bert 而言, 更加简单便捷、训练速度更快。

在中文词向量方面, 比较常用的方式是先对文章进行分词, 然后进行词向量训练。除了上述的一些词向量训练方法, 由于中文的特殊性, 研究者在针对中文的词向量训练上也进行了一些研究。Cao 等^[19]以 FastText 的 n -gram 的词向量训练为启发, 提出了基于笔画信息的 n -gram, 称之为 cw2vec 模型。Li 等^[20]和 Sun 等^[21]将中文字的部首信息融入词向量的生成过程中。Chen 等^[22]将组成词语的字抽取出来和词语一起进行训练, 提升了词向量生成的质量。Yu 等^[23]提出了汉字、字符和细粒度字字符的联合嵌入方法。在中文的词向量训练上的探索很多是在字符或更细粒度的汉字形态学上的语义特征提取, 如部首、笔画等及其结合。在词向量和文本表示中结合语言学的一些特征, 可以有效提升训练效率和训练效果, 如 Rei 等^[24]结合句子、词等不同粒度的监督目标, 进行更好的语言表达训练。

词向量作为一种词的数值化表示方法, 对于文本的数值化表示具有重要意义。一方面, 一些文本表示方法, 如词袋模型等, 需要依赖于词向量训练结果; 另一方面, 一些文本表示方法, 如 doc2vec 等, 其基本思想来源于基本的词向量训练原理。现有对于文本表示方法的研究, 主要集中于基于大型训练数据集的复杂网络、基于语言学模型和规律等方面, 并取得了很多有意义的研究进展。语言的复杂语义表示依然是具有挑战性的研究方向, 现有研究还具有很多不足之处。例如, Bert 等大型复杂网络依赖于大量的训练数据和强有力的硬件和训练时长要求; Word2vec 等常用的词向量训练模型在多义词的语义表示上有所欠缺; FastText 等训练模型仅集中于文本分类任

务, 可扩展性不强。

3 基于深度学习算法的文本分类技术

3.1 基于卷积神经网络的文本分类方法

卷积神经网络^[25]在计算机视觉等领域取得了不错的结果, 其卷积和池化的过程实际上是一个特征提取的过程。在数学上, 点积的几何意义是一个向量 u 在另一个向量 v 上的投影, 和 v 的长度相乘得到的值, 用数学表达式表示为 $uv = |u||v|\cos\theta$, 由于两个向量的夹角的 \cos 值是两个向量相似性的重要度量, 夹角越小表示两个向量所指的方向越接近, 从点积的公式可以看出, 点积是计算两个向量相似性的一种重要指标, 直观上, 形状越是相似的两个向量, 点积得到的值越大。卷积是一个信号倒序之后在另一个信号中不同位置同样大小的部分做了点积, 这个信号叫作卷积核, 点积越大的地方说明信号与卷积核的分布越相似, 因此, 卷积可以提取出与卷积核分布相似的信号区域, 于是起到了特征提取的作用。常用的卷积为宽卷积、窄卷积和等长卷积。卷积神经网络中通常会在卷积层后接池化层, 池化一般代表对特征响应图上给定区域求出一个能够代表这个区域特点的值, 能够在保留主要特征的同时减少下一层的参数和计算量、防止过拟合。在自然语言处理中最常用的池化方法为 Max-over-time 池化、k-Max 池化和 Chunk-Max 池化。总体来说, 卷积的作用是从句子中提取出局部的语义组合信息, 而多卷积核则是从多角度

对语义信息进行提取, 保证了提取语义组合的多样性; 池化的作用是对多种语义组合进行筛选, 过滤掉一些置信度低的组合^[26]。

Kim 等^[27]将 CNN 引入文本分类任务中, 一般称之为 TextCNN, 其网络结构如图 2 所示。文献^[27]中所用模型包括词向量层、卷积层、最大池化层、全连接层和 softmax 层共 5 层。模型输入为词向量组成的二维矩阵, 一行代表一个词的词向量, 卷积层采用宽度与词向量维度一致的卷积核, 以词为最小粒度提取特征, 高度可自定义, 这在一定程度上考虑了词序和上下文信息。池化层采用 Max-over-time 池化来提取最大特征, 最后接全连接层和 softmax 层进行分类操作。在卷积神经网络中有许多超参数要选择, 其中包括: 文本输入, 如 Word2vec、Glove、one-hot 等; 卷积核的数量、大小和类型, 卷积移动的步长; 池化策略, 如 k-max 池化、Average 池化等; 激活函数的选择, 如 ReLu、tanh 等。因此, Zhang 等^[28]针对文献^[27]基于 CNN 的文本分类模型在不同的超参数设置下进行了各种对比实验, 并给出了调参的建议和一些超参数的设置经验。

在基于 CNN 的文本分类神经网络模型上, 研究者一般围绕 Kim 等^[27]提出的基础模型, 在网络结构、网络深度、神经网络的训练速度和收敛等问题上进行改进、创新和研究。Wang 等^[29]先通过一个快速聚类算法在词向量矩阵进行语义特征提取, 再输入卷积层和 k-max 池化层进行分类。Kalchbrenner 等^[30]提出了一种名为 DCNN

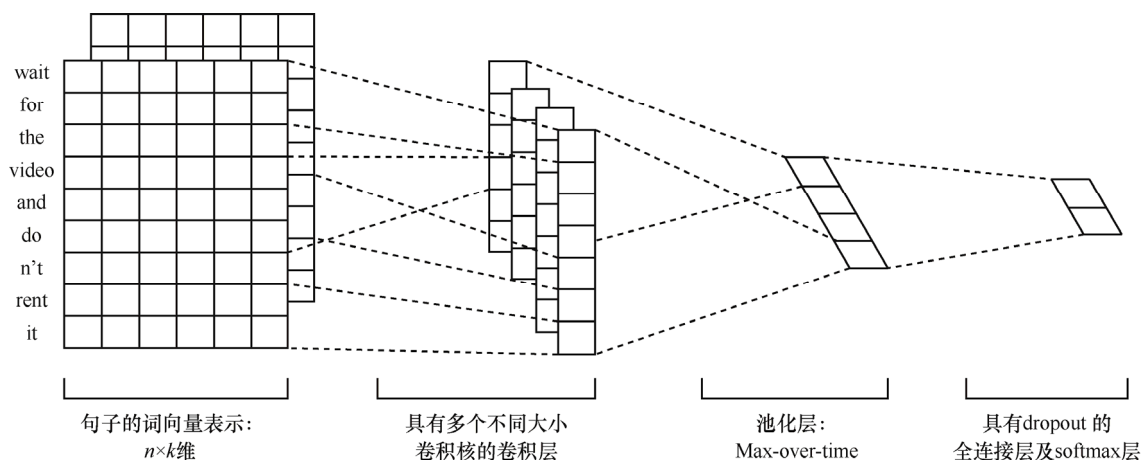


图2 TextCNN 结构
Figure 2 TextCNN structure

(dynamic convolutional neural network)的网络模型,采用多层卷积神经网络,能够处理可变长度的输入。DCNN 的卷积层采用宽卷积,即采用补零法在输入矩阵边缘添加 0 再进行卷积运算。在池化层,DCNN 提出了一种动态 k -max 池化的操作,动态 k -max 是指提取值最大的 k 个特征,其中 k 不是固定值,是输入句子长度和网络深度的函数,文献[30]给出的定义为 $K_l = \max\left(k_{\text{top}}, \frac{L-l}{L}s\right)$, 其

中 l 表示当前卷积层数, L 表示网络中卷积层总数, k_{top} 为固定值,表示最顶层卷积层对应的池化层的 k 值, s 为句子长度。从 K_l 定义可知, k 的大小随卷积神经网络的深度增加从底层开始逐渐衰减到顶层 k 的数值大小,即随着网络深度的加大,池化层保留的特征数量逐渐减少。此外,DCNN 使用折叠(folding)操作,即将两行的词向量相加,保留词向量矩阵中相邻两行之间的依赖。

以上均为浅层 CNN 应用于文本分类,一些研究已经证明浅层次的 CNN 在文本分类上能够取得不错的效果,但浅层次的 CNN 在捕捉文本长距离关联性上有所欠缺。研究发现,通过加深卷积层深度,可以捕捉到文本中的长距离依赖性。Conneau 等^[31]提出 VD-CNN 基于字符级别(character-level),利用小尺度卷积核与池化操作,通过只增加卷积层深度的方式证明在深度卷积神经网络下文本分类任务也可以取得很好的效果。腾讯 AI-lab 在 ACL2017 中提出的 DPCNN (deep pyramid convolutional neural networks)^[32]是一个 word-level 的广泛有效的深层文本分类卷积神经网络模型,并提出通过加深网络来提取文本中的长距离依赖关系。DPCNN 的底层保持了和 TextCNN 相似的结构,即包含多尺寸卷积核的卷积层(region embedding),不同点在于,TextCNN 采用保留词序的做法,而 DPCNN 采用的是词袋模型做法,即首先对 n -gram 的 n 个词的词向量取均值,然后用词向量长度的一维卷积进行卷积运算来提取特征。底层之上,DPCNN 采用等长卷积将输入序列每个词及其一定长度的上下文信息压缩成该词上一层的特征映射,在使每个词都得到一定上下文信息修饰的同时保持输出向量空间

与输入向量空间不变。卷积上层接 $\frac{1}{2}$ 池化层,将序列长度压缩至原来一半。这种等长卷积 + $\frac{1}{2}$ 池化层的方式使序列长度随网络层数加深而减小,上层序列长度的压缩,使上层的卷积层能够捕捉到更长距离上下文之间的依赖性。此外,为了解决深度 CNN 中的梯度弥散问题,文献[31]使用了 ResNet 中提出的残差连接(residual-connection)作为解决方案。通过加深网络深度,对于长文本的分类效果具有很大的提升。

除了在网络模型和深度等方面的研究和改进,还有针对文本输入的表现形式和训练状态的研究。在文本表现形式上,主要有不同方式的词向量表示,如 one-hot 或稠密词向量表示;不同的词向量训练方式得到的词向量,如随机初始化、第 2 节介绍的词向量训练方法得到的词向量等。在训练状态上,最常用的两种方法是固定不变的词向量表示和在训练同时对词向量矩阵进行优化的 fine-tuning 方式。Kim 等^[27]以不同的词向量输入以及训练过程中不同的词向量处理来进行对比实验;Zhang 等^[32]将字符级的文本当作原始输入来进行网络训练^[33];Johnson 等^[34]不以词向量作为网络输入,而是直接以 one-hot 表示输入网络进行卷积操作,还提出了一种空间高效的用于输入数据的单词表示法;Johnson 等^[35]以无监督的词向量训练过程为启发,提出了一种半监督的文本分类框架,具体来说,先用 CNN 从未标记数据中学习小文本区域的向量表示(tv-embedding),然后将其集成到后续的有监督训练网络中。另外,对于训练数据集中的噪声数据,Jindal 等^[36]引入了噪声模型和 CNN 一起训练以防止模型对噪声数据过拟合。

另外,由于文本表示形式的不同,一些研究将图卷积网络(GCN, graph convolutional network)^[37]引入文本分类任务(Text GCN, text graph convolutional network)。图卷积网络是一种多层神经网络,直接在图上运行,根据节点的领域属性来训练节点的向量表示。Text GCN 在图结构数据上运行,研究工作主要在于图构造方法的不同(如节点定义、边权重定义等)以及上层卷积网络的

不同。Kipf 等^[37]、Henaff 等^[38]、Defferrard 等^[39]、Peng 等^[40]、Yao 等^[41]均在基于图卷积网络的文本分类算法上有所研究和探索。

卷积神经网络通过卷积核提取文本的 n -gram 特征, 通过池化层进行特征选择, 通过加深网络深度捕捉文本的长距离依赖关系。图 3 总结了上文所列举的一些在文本分类任务中的 CNN 结构设计和常用的超参数调节手段。在卷积神经网络的网络参数选择和调整, 其网络结构的优化, 以及网络训练和运行速度的优化上, 研究人员进行了许多探索和研究, 其中, 很多研究成果对文本分类的工程实践和学术研究具有很大的借鉴和启发意义。

3.2 基于循环神经网络的文本分类技术

循环神经网络 (RNN, recurrent neural network) 以可变长度的序列数据作为数据输入, 其神经元具有自反馈功能, 是一类具有短期记忆能力的神经网络模型。RNN 的基本网络结构如图 4 所示, 其用数学表达式可以简单表示为 $x_t = F(x_{t-1}, u_t, \theta)$, 其中, x_t 表示时刻 t 的网络状

态, u_t 表示 t 时刻的网络输入。从图中可以看出, 循环神经网络为序列式输入的网络结构, t 时刻网络的状态和输出依赖于 $t-1$ 时刻的网络状态, 数据的时序信息在数据传递中得以保留。因此, 相对于前馈神经网络, RNN 在处理视频、语音、文本等时序数据上能力更强。将循环神经网络用于文本分类, 一般的策略是将文本序列通过循环神经网络进行特征提取映射为一个固定长度的文本向量表示, 然后将其输入 softmax 层进行分类^[42]。

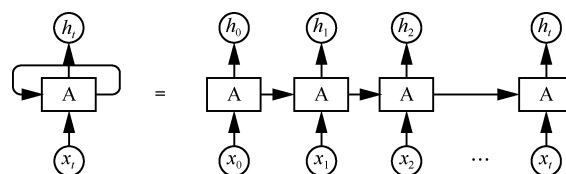


图4 RNN 结构
Figure 4 RNN structure

当输入序列比较长时, 简单的循环神经网络存在梯度爆炸和梯度消失问题, 虽然简单循环神经网络理论上可以建立长时间的文本依赖关系,

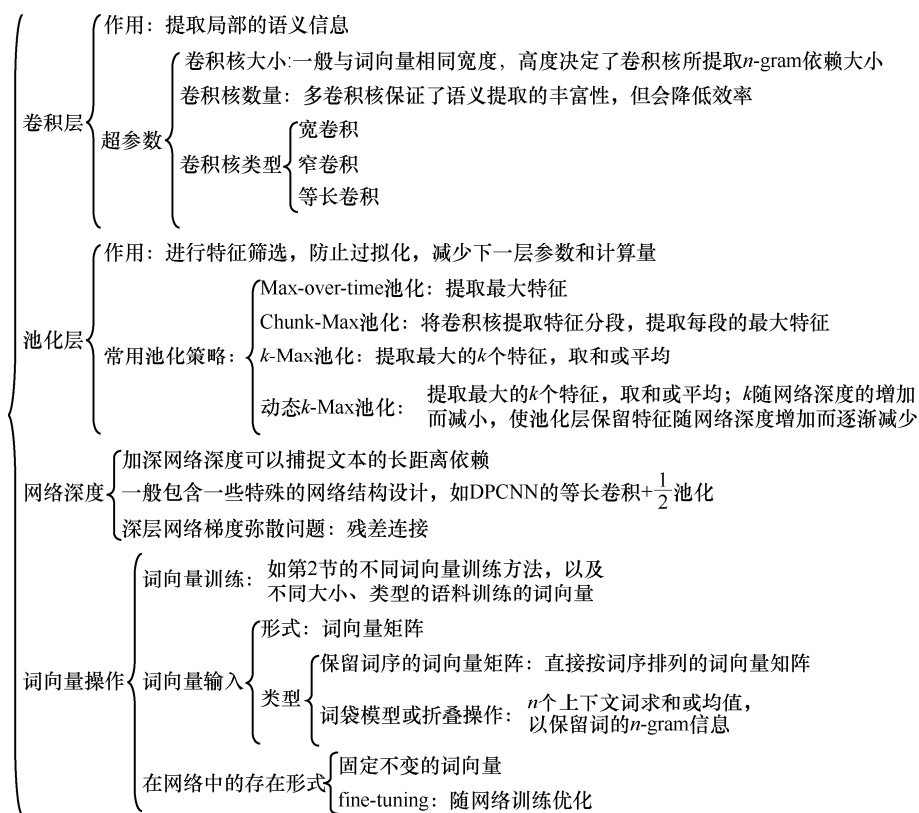


图3 文本分类中 CNN 超参数总结
Figure 3 The summary of CNN super parameter in text classification

但是由于梯度爆炸和梯度消失,实际上只能学习到短期的依赖关系,因此这个问题也被称为长期依赖问题^[43-45]。对于梯度爆炸问题,深度学习中一般采用权重衰减或梯度截断来避免。而对于梯度消失问题,优化模型是更有效的方式。其中,基于门控机制的循环神经网络是一种非常好的解决方案,其主要思想是有选择地遗忘以前的信息,有选择地加入新的信息。长短期记忆(LSTM, long short-term memory)网络^[44,46]和门控循环单元(GRU, gated recurrent unit)网络^[42,47]是两个经典的基于门控机制的循环神经网络。其中,标准的 LSTM 以遗忘门、输入门和输出门 3 个门来控制信息的遗忘和保存;GRU 简化了 LSTM 的网络结构,将 LSTM 的输入门与遗忘门合并成一个单一的更新门,GRU 和 LSTM 在许多任务上性能相似,但 GRU 结构上简化许多。

在一些任务中,为了同时获取文本的反向依赖关系,可以使用双向循环神经网络(Bi-RNN, bidirectional recurrent neural network)。Bi-RNN 由两层循环神经网络组成,其中一层按时间顺序传递,另一层按时间逆序传递,它可以同时获取文本的上下文依赖。LSTM 和 GRU 均有其对应的双向循环神经网络结构。文本循环神经网络(text RNN)模型的结构就是利用双向 LSTM 建模,使用最后一个单元输出向量作为文本特征连接 softmax 层分类。

RNN 在网络结构上产生了许多变体,如为解决梯度消失问题而提出的 LSTM、GRU 等基于门控机制的循环神经网络结构,为能够同时获取文本序列上下文依赖而提出的 Bi-RNN。在文本分类任务上,RNN 主要进行文本的特征提取而获得固定长度的文本表示,一般有以下几种策略:直接使用 RNN 最后一个单元输出向量作为文本特征;使用 Bi-RNN 两个方向的输出向量的连接或者均值作为文本特征。这两种策略受限于最后单元输出向量的文本表示,即使使用了门控机制的循环神经网络,在长文本序列下,梯度消失问题依旧存在。一种解决方案是将所有 RNN 单元的输出向量的均值池化或者最大池化作为文本特征;另一种有效的解决方案是注意力机制(attention mechanism)。

注意力机制早期用于图像领域,Google Mind 团队将注意力机制用于 RNN 进行图像分类取得了很好的结果^[48]。Bahdanau 等^[49]第一次使用类似于 Attention 的机制在机器翻译任务中,将其引入了 NLP 领域。注意力机制的基本思想是通过学习一系列的注意力分配系数,即权重系数,对 RNN 的每个单元输出向量进行权重分配,将其加权和作为文本特征。不同的注意力机制在权重系数的学习方式上有所不同。注意力模型的计算过程通常可描述为:将一个序列(query, Q)和一系列键值对(key-value pair, K-V)映射为一个输出的过程,其中这个输出通常表示为 value 值的加权和,而权重一般为 query 和 key 的函数。其一般计算方法分为 3 步:比较 Q 和 K 的相似度,表示为 $f(Q, K_i)$;将得到的相似度进行 softmax 归一

化得到权重, $\alpha_i = \frac{e^{f(Q, K_i)}}{\sum_{i=1}^n e^{f(Q, K_i)}}$;对 value 进行加权

求和,得到最后的 Attention 向量 $\sum_{i=1}^m \alpha_i V_i$ 。不同的

Attention 模型具有不同的相似度计算方法。Luong 等^[50]介绍了 Attention 机制如何在 RNN 中进行扩展,并提出了全局注意力(global attention)机制和局部注意力(local attention)机制。在全局注意力机制上,他们提出了几种简单的注意力系数,来计算方法的扩展版本。局部注意力机制主要思想是不考虑 RNN 的每个单元输出,而是根据一个预测函数得到当前需对其单元的输出位置 P,仅考虑 P 的上下文窗口内的单元输出。Yang 等^[51]提出的分层注意力网络(HAN, hierarchical attention networks)在文本分类任务中使用了注意力模型, HAN 的模型结构如图 5 所示。HAN 把文本结构分成词语、句子、文档 3 个层次,具有叠加在词语级别和句子级别的两层 Attention 机制。HAN 首先将每个句子的词序列的向量输入具有 Attention 机制的 Bi-GRU 网络得到每个句子的向量表达;然后将句子序列通过添加了 Attention 机制的 Bi-GRU 网络得到整个文本的向量表达;最后通过 softmax 层进行分类。HAN 的 Attention 机制实现方式是将输入序列在每个 Bi-GRU 单元的双向隐藏层输出进行连接,将其输入单层感知

机网络接 softmax 层得到这个单元输出的注意力分配系数, 这个模型随分类器一起训练。Google 在其提出的 Transformer 网络结构^[16]中使用了自注意力机制 (self-attention mechanism) 和编码器-解码器注意力机制 (encoder-decoder attention mechanism), 两者均是在 multi-head Attention 之上进行 softmax 归一化和加权求和的结果, 区别在于输入的不同。multi-head Attention 基于点积注意力 (scaled dot-product Attention) 模型, 是将其做 H 次后的输出合起来再经过一个线性层的结果。点积注意力模型的主要思想是通过向量点积计算相似度来获取权重。

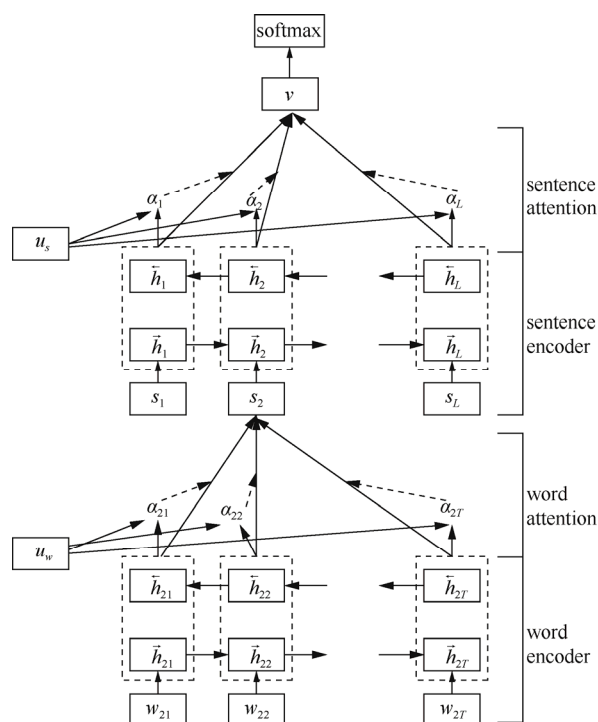


图5 HAN 结构
Figure 5 HAN structure

在不同的文本表示形式下, 相对于图结构, 递归神经网络 (RecNN, recursive neural network)^[52]是 RNN 在有向无循环图上的扩展, 一般为树状的层次结构。Socher 等^[53]将递归神经网络用于文本分类任务。Tai 等^[54]将树结构下的 LSTM 用于文本分类。

基于循环神经网络的文本分类方法, 一般通过 RNN 进行文本特征提取和文本向量化表示, 再用 softmax 层进行分类。RNN 以其结构上建立

的对上一时刻的网络状态的依赖来获取文本上下文间的依赖关系, 因此非常适用于具有时序性的语料。为了解决长时依赖问题, LSTM 和 GRU 等基于门控机制的循环神经网络以及适用于 RNN 的 Attention 机制被提出, 其中, 加上 Attention 机制的 RNN 模型能够取得更好的效果, 而 transformer 模型中只采用了精心设计的 Attention 机制, 而没有了 RNN 的结构。双向循环神经网络解决了单向 RNN 只能获取单向文本依赖的问题。在 RNN 结构中, seq2seq 的网络结构和 Attention 模型的提出具有里程碑式的意义, 极大地促进了深度学习在自然语言处理领域的发展。但是, 相对于 CNN, RNN 结构的数学理论依据较为薄弱, 其更倾向于在语言学上的理解, 这也在一定程度上限制了对其更为深入的研究。

3.3 其他基于深度学习的文本分类技术

卷积神经网络和循环神经网络是在文本分类中最常用的两种网络结构。在基于卷积神经网络的文本分类方面, 一般采用宽度与词向量维度一致的卷积核进行特征提取, 以池化层进行特征选择, 通过加深网络深度捕捉文本的长距离依赖, 并通过 softmax 层进行分类。在基于循环神经网络的文本分类方面, 一般基于 seq2seq 中的编码器结构进行文本特征提取, 将其表示为固定长度的文本特征向量, 再输入 softmax 层进行分类。为了进行更好的文本特征提取以利于分类, 基于门控机制的 RNN、双向 RNN 和 Attention 机制是有效的解决方案。卷积神经网络和循环神经网络各有其优缺点, 一些工作将 CNN 和 RNN 相结合用于文本分类。Lai 等提出的 RCNN (recurrent convolutional neural networks) 模型^[55]利用双向 RNN 获取单词的上文特征表示和下文特征表示, 将其与这个词本身的词向量连接经过一个线性层来代替 CNN 中的卷积核特征提取部分, 并将其输入最大池化层和 softmax 层进行分类。Zhou 等^[56]提出的 C-LSTM 网络先使用卷积层进行特征提取, 将同一位置不同卷积核的特征提取结果合在一起, 作为这个词的特征向量表示, 将其输入 LSTM 网络进行分类。Tang 等^[57]提出了一种具有层次结构的文本分类结构, 先用 LSTM 或者 CNN 对生成句子表示,

用一种基于门控的 RNN 网络进行特征提取生成文档向量,再输入 softmax 进行分类。Xu 等^[58]提出了一种多通道 RNN 模型 (MC-RNN),进行动态捕获动态语义结构信息。

通过 RNN 或 CNN 进行文本分类,可以达到比较好的效果,但是也存在训练时间长、超参数调整麻烦等问题。针对这些问题,一些基于简单模型的文本分类方法被提出。DAN^[59]以文本中每个词的词向量的求和平均作为神经网络输入,叠加几个线性的隐藏层,再输入 softmax 层分类。DAN 还提出了一种 word dropout 策略,即在求词向量的平均值之前,随机使文本中某些词失效。Facebook 发布的 FastText 快速文本分类工具^[18]也采用以文本的词向量均值作 softmax 进行分类的策略,但 FastText 以所有的词及其词级别的 n -gram 向量共同的叠加平均作为文档向量。并且,在分类类别较多时,可采用层次 softmax 提高训练速度。FastText 在大型训练集下的分类效果与基于 CNN 和 RNN 的分类模型相差无几,但大大缩短了训练时间。Adhikari 等^[60]提出简单的 Bi-Lstm 模型通过正则等参数设置也能达到良好的文本分类效果。

4 已有的研究工作总结和未来研究重点

词向量作为文本在词粒度上的数值化表示形式,其研究和发展对文本分类算法的发展具有重要意义。词向量的发展经历了从独热编码到稠密化分布式向量表示的发展历程,分布式词向量一般是语言模型训练过程中的产物,可以通过欧几里得距离衡量词语之间的相似性,大大缓解了维度灾难的问题。目前常用于词向量训练的语言模型为 CBOW 和 Skip-gram,其基本思想均基于上下文预测,但是这种方法获得的词向量无法表示多义词语义。为解决单一词向量无法表示多义词的问题,ELMo、GPT、Bert 等更为复杂的模型被提出,但其训练时间长、硬件要求高且对后续的自然语言处理任务进行有一定的流程要求,易用性不高,不能像 CBOW 和 Skip-gram 模型得到的词向量一样随取随用。在中文词向量上,研究者在汉字形态学上的语义提取有所研究。现有对于词向量的研究大多基于上下文预测,结合语言学

的相关知识,在特征工程、模型结构、训练方式等方面进行研究。现有研究的不足在于:第一,对于词向量质量尚未有较为完善的统一的衡量标准;第二,对词的多语义数值化表示的研究仍处于起步阶段,现有研究成果对于训练成本和训练质量难以两全,或者比较专注于一项自然语言处理任务或某一领域,缺少可扩展性;第三,所获的语言数值化表示对于后续自然语言处理任务的易用性。文本和语言的数值化表示是一切自然语言处理任务的基础,而词向量的丰富语义表示主要依赖于两点:多义词语言信息的数值化表示和训练方法;高质量的大型训练数据集。基于这两点,在词向量训练模型中,词向量抽取方式、词向量的连续性学习等方面是重要的研究方向。

现有的基于深度学习的文本分类算法,其基本过程是利用 RNN 或 CNN 等网络进行文本特征提取和文本的向量化表示,然后输入 softmax 层进行文本分类。卷积神经网络和循环神经网络是两个常用于文本分类算法的网络结构。卷积神经网络以点积相似性为数学基础,提取与卷积核相似度较高的文本特征,多卷积核保证了特征提取的多样性,以池化层进行特征选择和降维、防止过拟合,研究发现加深 CNN 深度可以提取到文本的长距离关联。对于基于卷积神经网络的文本分类算法,研究者在词向量表示,卷积核数量、大小、步长,池化策略,激活函数选择,网络深度等方面有很多有意义的探索和研究成果。循环神经网络以编码器结构为基本的文本特征提取结构,非常适用于序列化的文本数据。为解决简单循环神经网络的长时依赖问题,LSTM 和 GRU 等基于门控机制的循环神经网络被提出;针对同时获取词的上下文依赖,Bi-RNN 是不错的解决方案,LSTM 和 GRU 都有其对应的双向结构;Attention 机制是另一有效解决长时依赖问题的方案,基于 Attention 机制的 RNN 已成为主流。很多研究者将 RNN 与 CNN 相结合用于文本分类,也取得了不错的结果。基于 CNN 和 RNN 的文本分类算法,在大型文本数据训练集下有很长的训练时间,有些还需要依赖于强大的计算机硬件。DAN 和 FastText 等简单网络分类算法被提出,且在一些大型的数据集上取得了不

错的效果,大大缩短了训练时间。对于深度学习文本分类算法的研究,主要集中在特征工程、网络模型设计、结合语言学特征的训练方法等方面。总体上呈现出两种方向的思想:通过加深网络和设计更复杂的网络结构来实现丰富的语义表达和获取长距离语义间的关联信息;通过简单的网络结构和大量的数据集训练减少训练成本,但依旧能保持不错的训练效果。

文本分类算法一直是自然语言处理的一个重要研究方向,时至今日,研究者已经提出了许多影响深远的文本分类方法,这些方法有些已经投入工业应用中,有些对后续的研究有很大的启发意义。未来,仍有很多问题需要进一步的研究和探索,以在算法的准确率、训练和运行效率等方面取得更好的效果。在基于深度学习的文本分类算法方面的研究和探索主要体现在以下几个方面。

1) 对词向量来说,有效和统一的词向量质量衡量标准,以及高质量的中文词向量公开数据集,对于后续文本分类等自然语言处理任务的进行起到了很大的作用;使用词向量公开数据集也更便于对文本分类算法等神经网络模型的有效性进行更客观的模型对比和评价。

2) 在词向量训练方面,更为高效地考虑多义词情景下的词向量训练和表示方法,以及相对 Bert 等需要将分类网络整合进预训练网络结构进行后续任务来说,自由度更大的词向量取用方式,也是一个有意义的研究方向。另外,因为词语义的多样性,词向量的可持续学习也是一个重要的研究方向。

3) 在基于深度神经网络的文本分类算法方面,对不同文本分类网络结构的探索和研究仍是研究热点。一方面,研究者将已提出的神经网络模型和机制,如 RNN、CNN、Attention 机制、不同的 RNN 和 CNN 变体等进行基于一定理论基础的有机结合,组成新的网络结构进行实验和研究。例如,由于 Attention 机制取得的巨大成功,且 CNN 在文本分类上也取得了不错的效果,将 Attention 机制与 CNN 相结合,是目前的一个不错的研究方向。另一方面,一些研究者专注于网络结构本身,通过创新、理论研究和公式推算,提出新的网络结构,如 CNN、LSTM、GRU、Attention 机制等的初次提出,这些新的思想和

探索对于深度学习的发展具有很大的启发意义,如新的 Attention 权重计算方法、新的特征提取网络结构等都是现在的研究热点。目前,对于很多神经网络结构的理论基础依旧较为薄弱,这也是具有创新性的网络模型结构难以提出及发展缓慢的重要原因。

4) 对于深度神经网络的训练成本消耗问题,以及神经网络训练过程中梯度消失等问题的解决方案也是一个重要的研究方向。由于深度学习技术立足于大数据,在一些研究和训练任务中,需要大量的 CPU 和 GPU 资源,耗费很长的训练时间,此时,有效的优化手段可以节省大量的训练成本消耗,加快训练速度。在一些网络深度较大或计算比较复杂的神经网络结构中,可能会出现梯度消失、梯度爆炸等问题。对这些问题,需要研究者深入分析原因,找到其弊端所在,提出新的解决方案。

5 结束语

本文主要从词向量的发展和研究现状、基于深度学习的文本分类方法的发展和研究现状两方面介绍了在深度学习算法高度发展的背景下,人们将其应用于文本分类领域的思考和研究。近年来,在文本的表示方法和文本分类方法领域,有许多具有重要意义的研究成果,如 Attention 机制、FastText、Bert 等网络模型的提出。但是,深度学习在自然语言处理领域的研究才刚刚起步,其长足发展还需依赖于研究者对语言学、神经网络等方面深入的研究和结合。

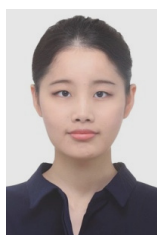
参考文献:

- [1] 于游,付钰,吴晓平. 中文文本分类方法综述[J]. 网络与信息安全学报, 2019, 5(5): 1-8.
YU Y, FU Y, WU X P. Summary of text classification methods[J]. Chinese Journal of Network and Information Security, 2019, 5(5): 1-8.
- [2] 明拓思宇,陈鸿昶. 文本摘要研究进展与趋势[J]. 网络与信息安全学报, 2018, 4(6): 1-10.
MING T S Y, CHEN H C. Research progress and trend of text summarization[J]. Chinese Journal of Network and Information Security, 2018, 4(6): 1-10.
- [3] HINTON G E. Learning distributed representations of concepts[C]//Proceedings of the Eighth Annual Conference of the Cognitive Science Society. 1986: 12.

- [4] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model[J]. *Journal of Machine Learning Research*, 2003, 3(2): 1137-1155.
- [5] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. *Journal of Machine Learning Research*, 2011, 12(8): 2493-2537.
- [6] MNIH A, HINTON G. Three new graphical models for statistical language modelling[C]//*Proceedings of the 24th International Conference on Machine Learning*. 2007: 641-648.
- [7] MNIH A, HINTON G E. A scalable hierarchical distributed language model[C]//*Advances in Neural Information Processing Systems*. 2009: 1081-1088.
- [8] TOMÁŠ M. Statistical language models based on neural networks[D]. Brno: Brno University of Technology, 2012.
- [9] MIKOLOV T, CHEN K, CORRADO G, et al. Estimation of word representations in vector space[C]//*International Conference on Learning Representations*. 2013.
- [10] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//*Advances in Neural Information Processing Systems*. 2013: 3111-3119.
- [11] MNIH A, KAVUKCUOGLU K. Learning word embeddings efficiently with noise-contrastive estimation[C]//*Advances in Neural Information Processing Systems*. 2013: 2265-2273.
- [12] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]//*Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014: 1532-1543.
- [13] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword information[J]. *Transactions of the Association for Computational Linguistics*, 2017, 5: 135-146.
- [14] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[J]. *arXiv preprint arXiv:1802.05365*, 2018.
- [15] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[R]. 2018.
- [16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//*Advances in Neural Information Processing Systems*. 2017: 5998-6008.
- [17] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[J]. *arXiv preprint arXiv:1810.04805*, 2018.
- [18] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification[J]. *arXiv preprint arXiv:1607.01759*, 2016.
- [19] CAO S, LU W, ZHOU J, et al. cw2vec: learning chinese word embeddings with stroke n-gram information[C]//*Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.
- [20] LI Y, LI W, SUN F, et al. Component-enhanced Chinese character embeddings[C]//*Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015: 829-834.
- [21] SUN Y, LIN L, YANG N, et al. Radical-enhanced chinese character embedding[C]//*International Conference on Neural Information Processing*. Springer, Cham, 2014: 279-286.
- [22] CHEN X, XU L, LIU Z, et al. Joint learning of character and word embeddings[C]//*Twenty-Fourth International Joint Conference on Artificial Intelligence*. 2015.
- [23] YU J, JIAN X, XIN H, et al. Joint embeddings of chinese words, characters, and fine-grained subcharacter components[C]//*Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017: 286-291.
- [24] REI M, SØGAARD A. Jointly learning to label sentences and tokens[C]//*Proceedings of the AAAI Conference on Artificial Intelligence*. 2019, 33: 6916-6923.
- [25] LECUN Y, BOSER B, DENKER J S, et al. Backpropagation applied to handwritten zip code recognition[J]. *Neural Computation*, 1989, 1(4): 541-551.
- [26] HU B, LU Z, LI H, et al. Convolutional neural network architectures for matching natural language sentences[C]//*Advances in Neural Information Processing Systems*. 2014: 2042-2050.
- [27] KIM Y. Convolutional neural networks for sentence classification[C]//*Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014: 1746-1751.
- [28] ZHANG Y, WALLACE B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification[J]. *arXiv preprint arXiv:1510.03820*, 2015.
- [29] WANG P, XU J, XU B, et al. Semantic clustering and convolutional neural network for short text categorization[C]//*Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. 2015: 352-357.
- [30] KALCHBRENNER N, GREFFENSTETTE E, BLUNSOM P. A convolutional neural network for modelling sentences[J]. *arXiv preprint arXiv:1404.2188*, 2014.
- [31] CONNEAU A, SCHWENK H, BARRAULT L, et al. Very deep convolutional networks for text classification[J]. *arXiv preprint arXiv:1606.01781*, 2016.
- [32] JOHNSON R, ZHANG T. Deep pyramid convolutional neural networks for text categorization[C]//*Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017: 562-570.
- [33] ZHANG X, ZHAO J, LECUN Y. Character-level convolutional networks for text classification[C]//*Advances in Neural Information Processing Systems*. 2015: 649-657.
- [34] JOHNSON R, ZHANG T. Effective use of word order for text categorization with convolutional neural networks[J]. *arXiv preprint arXiv:1412.1058*, 2014.
- [35] JOHNSON R, ZHANG T. Semi-supervised convolutional neural networks for text categorization via region embedding[C]//*Advances in Neural Information Processing Systems*. 2015: 919-927.
- [36] JINDAL I, PRESSEL D, LESTER B, et al. An effective label noise model for dnn text classification[J]. *arXiv preprint arXiv:1903.07507*, 2019.
- [37] KIPF T N, WELING M. Semi-supervised classification with graph convolutional networks[J]. *arXiv preprint arXiv:1609.02907*, 2016.
- [38] HENAFF M, BRUNA J, LE CUN Y. Deep convolutional networks on graph-structured data[J]. *arXiv preprint arXiv:1506.05163*, 2015.

- [39] DEFFERRARD M, BRESSON X, VANDERGHEYNST P. Convolutional neural networks on graphs with fast localized spectral filtering[C]//Advances in Neural Information Processing Systems. 2016: 3844-3852.
- [40] PENG H, LI J, HE Y, et al. Large-scale hierarchical text classification with recursively regularized deep graph-CNN[C]//Proceedings of the 2018 World Wide Web Conference. 2018: 1063-1072.
- [41] YAO L, MAO C, LUO Y. Graph convolutional networks for text classification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 7370-7377.
- [42] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. Computer Science, 2014.
- [43] BENGIO Y, SIMARD P, FRASCONI P. Learning long-term dependencies with gradient descent is difficult[J]. IEEE Transactions on Neural Networks, 1994, 5(2):157-166.
- [44] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
- [45] HOCHREITER S, BENGIO Y, FRASCONI P, et al. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies[M]//A Field Guide to Dynamical Recurrent Networks, 2001: 237-243.
- [46] GERS F A, SCHMIDHUBER J, CUMMINS F. Learning to forget: continual prediction with LSTM[C]//International Conference on Artificial Neural Networks. 2002.
- [47] CHUNG J, GULCEHRE C, CHO K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv preprint arXiv:1412.3555, 2014.
- [48] MNH V, HEES N, GRAVES A, et al. Recurrent models of visual Attention[C]//Advances in Neural Information Processing Systems. 2014.
- [49] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.
- [50] LUONG M T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation[J]. arXiv preprint arXiv:1508.04025, 2015.
- [51] YANG Z, YANG D, DYER C, et al. Hierarchical attention networks for document classification[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 1480-1489.
- [52] POLLACK J B. Recursive distributed representations[J]. Artificial Intelligence, 1990, 46(1):77-105.
- [53] SOCHER R, PERELGIN A, WU J Y, et al. Recursive deep models for semantic compositionality over a sentiment Treebank[C]//Proc. EMNLP. 2013.
- [54] TAI K S, SOCHER R, MANNING C D. Improved semantic representations from tree-structured long short-term memory networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015: 1556-1566.
- [55] LAI S, XU L, LIU K, et al. Recurrent convolutional neural networks for text classification[C]//Twenty-ninth AAAI Conference on Artificial Intelligence. 2015.
- [56] ZHOU C, SUN C, LIU Z, et al. A C-LSTM neural network for text classification[J]. arXiv preprint arXiv:1511.08630, 2015.
- [57] TANG D, QIN B, LIU T. Document modeling with gated recurrent neural network for sentiment classification[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 1422-1432.
- [58] XU C, HUANG W, WANG H, et al. Modeling local dependence in natural language with multi-channel recurrent neural networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 5525-5532.
- [59] IYYER M, MANJUNATHA V, BOYD-GRABER J, et al. Deep unordered composition rivals syntactic methods for text classification[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015, 1: 1681-1691.
- [60] ADHIKARI A, RAM A, TANG R, et al. Rethinking complex neural network architectures for document classification[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019: 4046-4051.

[作者简介]



杜思佳(1995-), 女, 浙江嘉兴人, 哈尔滨工业大学硕士生, 主要研究方向为网络舆情分析、网络安全。



于海宁(1983-), 男, 黑龙江鹤岗人, 博士, 哈尔滨工业大学助理研究员, 主要研究方向为物联网安全搜索与隐私保护、云安全与隐私保护。



张宏莉(1973-), 女, 吉林榆树人, 博士, 哈尔滨工业大学教授、博士生导师, 主要研究方向为网络与信息安全、网络测量与建模、网络计算、并行处理。