

文章编号: 1003-0077(2021)02-0069-09

基于层次混合注意力机制的文本分类模型

孙 新, 唐 正, 赵永妍, 张颖捷

(北京市海量语言信息处理与云计算应用工程技术研究中心 北京理工大学 计算机学院, 北京 100081)

摘 要: 文本分类是自然语言处理领域的核心任务之一, 深度学习的发展给文本分类带来更广阔的发展前景。针对当前基于深度学习的文本分类方法在长文本分类中的优势和不足, 该文提出一种文本分类模型, 在层次模型基础上引入混合注意力机制来关注文本中的重要部分。首先, 按照文档的层次结构分别对句子和文档进行编码; 其次, 在每个层级分别使用注意力机制。句编码时在全局目标向量基础上同时利用最大池化提取句子特定的目标向量, 使编码出的文档向量具有更加明显的类别特征, 能够更好地关注到每个文本最具区别性的语义特征。最后, 根据构建的文档表示对文档分类。在公开数据集和行业数据集上的实验结果表明, 该模型对具有层次结构的长文本具有更优的分类性能。

关键词: 文本分类; 深度学习; 注意力机制

中图分类号: TP391 文献标识码: A

Hierarchical Networks with Mixed Attention for Text Classification

SUN Xin, TANG Zheng, ZHAO Yongyan, ZHANG Yingjie

(Beijing Engineering Applications Research Center on High Volume Language Information Processing and Cloud Computing, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

Abstract: Text classification is one of the core tasks in the field of natural language processing. To address the long text sequence, we propose the Hierarchical Networks with Mixed Attention (HMAN) model for text classification to capture the important parts of the text based on the hierarchical model. First, the sentences and documents are encoded according to the hierarchical structure of documents, and attention mechanism is applied at each level. Then, global target vectors, sentence specific target vectors are extracted by max-pooling to encode the document vectors. Finally, documents are classified according to the constructed document representation. Experimental results on the open datasets and industry text datasets show that the model has better classification performance, especially for long text with hierarchical structure.

Keywords: text classification; deep learning; attention mechanism

0 引言

文本分类是自然语言处理的核心任务之一, 在新闻分类、情感分析、主题检测等领域均有广泛应用。传统的机器学习分类算法通过特征工程选择合适的特征来表示文本, 然后把文本特征输入到不同的分类模型中得到分类结果, 例如, 朴素贝叶斯 (Naïve Bayes, NB)^[1]、K 近邻 (K-nearest neighbor,

KNN)^[2]、支持向量机 (support vector machine, SVM)^[3] 等算法。但是, 随着文本数量的日益庞大, 传统机器学习的分类算法中复杂的特征工程成了制约其发展的瓶颈。

深度学习模型能够从大规模数据样本中自主学习样本的特征, 提高了建模的智能化, 简化了分类流程, 因此成为文本分类领域的研究热点。卷积神经网络 (convolutional neural network, CNN)^[4] 是深度学习中的常用模型之一。纽约大学的 Yoon Kim 在 2014

收稿日期: 2019-12-18 定稿日期: 2020-02-24

基金项目: 国家重点研发计划项目 (2017YFB0803300)

年的 EMNLP 会议^[4]中提出了 TextCNN 模型^[5],利用卷积神经网络对文本建模并分类,得到了不逊色于复杂的基于机器学习的分类器模型的结果,由此引发了对基于深度学习的文本分类模型研究的热潮。

在自然语言处理任务中,循环神经网络(recurrent neural network, RNN)通常一次处理一个单词,并根据复杂的单词序列学习特征,因此 RNN 能够捕获对自然语言处理任务有用的语言模式,尤其是在较长的文本段上。Liu 等人^[6]利用基于长短期记忆(LSTM)的循环神经网络对文本进行编码获得了较好的效果。

卷积神经网络可以提取局部特征,循环神经网络可以提取全局特征,均表现出不错的效果。但是,当文档长度较长时,直接把文档作为长序列处理,不仅会给模型的性能带来很大挑战,同时也会忽略掉文档层次结构中包含的信息。因此有研究者研究分层的神经网络模型^[7-9],使用层次网络模型进行文本分类,但是层次网络模型在训练过程中通常使用全局目标向量,无法关注到每个文本最明显的语义特征。

在对句子使用注意力机制时,当前已有方法通常使用全局参数作为所有类别的目标向量,这一方面不利于表示每个句子各自的特征,另一方面也不能突出句子中具有明显类别特征的词。针对当前文本分类算法在长文本分类中的优势和不足,本文提出基于层次混合注意力机制的文本分类模型(hierarchical mixed attention networks, HMAN),采用基于 RNN 的层次模型对长文本分类,同时引入注意力机制来关注文本中的重要部分。在句注意力层,除了训练一个全局目标向量之外,还利用最大池化从句子的词向量矩阵中提取每个维度上最重要的信息作为句子特定的目标向量,使用两个目标向量共同对句子中的单词打分,从而使得到的句子编码的类别特征更加明显,能够更好地关注到每个文本最具区别性的语义特征。

最后,在公开数据集以及企业年报数据集上进行实验,验证了 HMAN 模型的有效性。尤其针对具有层次特征、长文本的企业年报数据集, HMAN 模型在句注意力层通过提取句子中各维度的最大特征来获取每个句子特定的重要特征,在一级行业分类和二级行业分类的分类准确率上均有良好表现。

1 相关工作

基于深度学习的文本分类算法通常使用低维、实数值的词向量来表示文本中的单词,然后构建神

经网络模型对文本建模,获得包含了全部文本信息的文本表示,用最终的文本表示来对文本分类。

TextCNN 模型^[5]是经典的基于深度学习的文本分类模型。之后,研究者陆续提出了许多基于 TextCNN 的改进方案。Xiao 等人^[10]提出了基于 CNN 的字符级别的分类模型,在 CNN 的基础上加上了一层循环层来捕获句子中长期依赖的信息。Conneau 等人^[11]则是通过增加 CNN 的深度来获取序列上的依赖信息。Johnson 和 Zhang^[12]研究了如何加深 CNN 的词粒度对文本进行全局表达,提出一种简单的金字塔型的 CNN 网络结构,既增加了网络深度、提升了准确率,又没有过多地增加计算量。

受到 TextCNN 的启发,并考虑到 RNN 在处理文本数据时的优势,也有研究者使用 RNN 及其变体结构对文本进行建模并分类。Liu 等人^[6]利用 LSTM 结构对文本进行编码,同时为了解决单个任务中标注数据较少的问题,基于 RNN 设计了三种不同的信息共享机制进行训练,并在四个基准的文本分类任务中都获得了较好的效果。为了解决 RNN 在对长文本进行编码时存储单元不足的问题, Xu 等人^[13]提出了一种具有高速缓存的 LSTM 结构来捕获长文本中的整体语义信息,从而使网络能够在一个循环单元中更好地保存情感信息。

除了能够捕捉长文本上的序列信息, RNN 用于文本分类的另一个优势在于其可以很好地与注意力机制(attention mechanism)结合,在文本建模时把注意力重点放在关键信息上,从而提高文本分类的效果。注意力机制最初是在计算机视觉领域提出的^[14],在自然语言处理领域,注意力机制最先被引入到机器翻译任务的基于 RNN 的编解码器模型中^[15]。注意力机制通过目标向量对输入序列打分,把注意力集中在输入序列中更重要的部分,使输出结果更加精确,因而逐渐被推广应用到包括文本分类在内的多种的 NLP 任务中。

在处理由许多句子组成的长文档表示时,直接把文档作为长序列处理,会忽略掉文档的层次结构中包含的信息,因此有研究者采用分层的神经网络模型对文档建模来进行文本分类。Tang 等人^[7]构建了自底向上的文档表示方法,先用 CNN 对句子进行编码,然后利用带有门控结构的 RNN 构建文档表示,最后通过 softmax 层得到分类结果,实验证明,这种模型在当时得到了对长文档分类的最好效果。类似地, Yang 等人^[8]提出了分层注意力模型,

该模型将注意力机制纳入分层 GRU 模型,使模型能更好地捕获文档的重要信息,进一步提高了长文档分类的准确率。

2 基于层次混合注意力机制的文本分类模型

为了捕捉长文本上的序列信息,同时更好地利用长文档数据中的层次结构,本文提出基于层次混合注意力机制的文本分类模型(hierarchical mixed attention networks, HMAN),使用基于 RNN 的层次模型,引入注意力机制关注文本中的重要部分,在句注意力层通过提取句子中各维度的最大特征来获取每个句子的重要特征。

模型的基本思想是:按照单词组成句子、句子组成文档的层次结构分别对句子和文档编码。为了把句子或文档表示的语义重点放在其重要的组成成分上,在每个层级分别使用注意力机制。最后根据构建的文档表示对文档分类。

通常注意力机制首先设定一个任务特定的目标

向量并与输入序列匹配,然后通过计算输入序列中每个元素与目标向量的相似度来为其分配注意力得分,将得分归一化后,对所有元素加权求和得到最终的表示结果。在文本分类中,以往的做法是在网络中学习一个全局的上下文向量作为目标向量,通过计算每个单词与目标向量的相似度对单词打分。

然而,当所有类别共同使用一个目标向量时,它在每个特征维度上的信息就会相对平均,不能突出句子的显著特征。即使句子里出现了具有明显类别特征的词,全局的目标向量也无法为它分配一个与其显著性相匹配的注意力得分。

因此,HMAN 模型在句注意力层使用混合的注意力机制,即除了使用全局目标向量之外,对每个句子构建其特有的目标向量。直接从句子的词向量矩阵中抽取每个维度上最大的值,也就是最明显的信息作为句子特有的目标向量。

HMAN 模型共有五层,自底向上分别为:句编码层、句注意力层、文档编码层和文档注意力层和文档分类层,模型结构如图 1 所示。

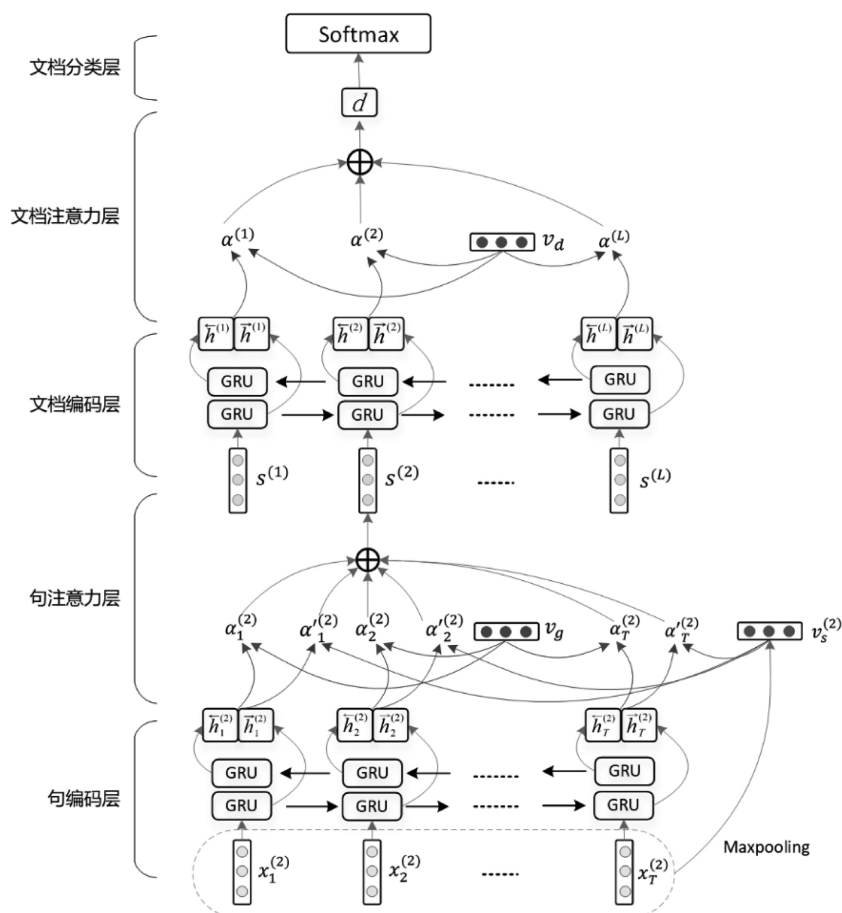


图 1 基于层次混合注意力机制的文本分类模型

2.1 句编码层

在句编码层,将句子中的每个词依次输入来构建句子表示。为了获取句子中的序列信息,使用 RNN 对句子建模。然而在基础的 RNN 中,随着序列在时间上的传播,序列中的历史信息会逐渐被遗忘,而误差累积却越来越多。因此在 HMAN 中,使用特殊的 RNN 结构——门循环单元 (gate recurrent unit, GRU) 解决长期记忆和反向传播中的梯度更新问题。

对于一个句子 $s^{(i)} = [x_1^{(i)}, x_2^{(i)}, \dots, x_T^{(i)}]$, $x_t^{(i)}$ 是句子 $s^{(i)}$ 中的第 t 个词的向量表示, $t \in [1, T]$ 。使用双向 GRU,通过汇总来自两个方向的单词信息获得结合了上下文信息的单词注解,如式(1)、式(2)所示。

$$\vec{h}_t^{(i)} = \overrightarrow{\text{GRU}}(x_t^{(i)}), t \in [1, T] \quad (1)$$

$$\overleftarrow{h}_t^{(i)} = \overleftarrow{\text{GRU}}(x_t^{(i)}), t \in [1, T] \quad (2)$$

其中, $\overrightarrow{\text{GRU}}$ 和 $\overleftarrow{\text{GRU}}$ 代表两个方向上的 GRU 结构,分别从前向后和从后向前进行编码。对于每个单词 $x_t^{(i)}$,连接两个方向上的隐藏层状态 $\vec{h}_t^{(i)}$ 和 $\overleftarrow{h}_t^{(i)}$ 作为它的注解 $h_t^{(i)}$,这个注解表示以 $x_t^{(i)}$ 为中心的整个句子信息,如式(3)所示。

$$h_t^{(i)} = [\vec{h}_t^{(i)}, \overleftarrow{h}_t^{(i)}] \quad (3)$$

2.2 句注意力层

由于句子中每个单词对分类目标的贡献度是不一样的,在编码时对分类越重要的词的权值应该越大。因此在句注意力层,采用注意力机制给句子中的每个单词计算一个注意力得分,然后通过单词及其得分形成句子的向量表示。

特别地,在句注意力层,HMAN 模型使用混合的注意力机制,除了使用全局目标向量 v_g 之外,还为每个句子构建其特有的目标向量 v_s 。由于词向量的每个维度都表示了一个属性信息,类似于 CNN 中的最大池化 (Maxpooling),直接从句子的词向量矩阵中抽取每个维度上最大的值,也就是最明显的信息作为句子特有的目标向量,更加突出具有明显类别特征的语义信息。

对于文档中的每个句子, $s^{(i)} = [x_1^{(i)}, x_2^{(i)}, \dots, x_T^{(i)}]$,其中 $x_t^{(i)}$ 是句子 $s^{(i)}$ 中的第 t 个单词的向量表示, $t \in [1, T]$ 。每个单词的词向量 $x_t^{(i)} = [x_{t1}^{(i)}, x_{t2}^{(i)}, \dots, x_{tW}^{(i)}]$, W 是短语向量维度。

计算句子特定的目标向量 $v_s^{(i)}$,结合全局目标向量 v_g ,按照一定的权重相加作为句子的注意力得

分,得到句子的最终编码值。如图 2 所示,在计算句子特定的目标向量 $v_s^{(i)}$ 时,对全部 T 个单词的 W 个维度,取每个维度上的最大值作为特征,然后将全部 W 个维度上的最大值连接起来作为句子 $s^{(i)}$ 特有的目标向量 $v_s^{(i)}$,如式(4)、式(5)所示。

$$v_s^{(i)} = [u_1^{(i)}, u_2^{(i)}, \dots, u_W^{(i)}] \quad (4)$$

$$u_j^{(i)} = \max(x_{1j}^{(i)}, x_{2j}^{(i)}, \dots, x_{Tj}^{(i)}), j \in [1, W] \quad (5)$$

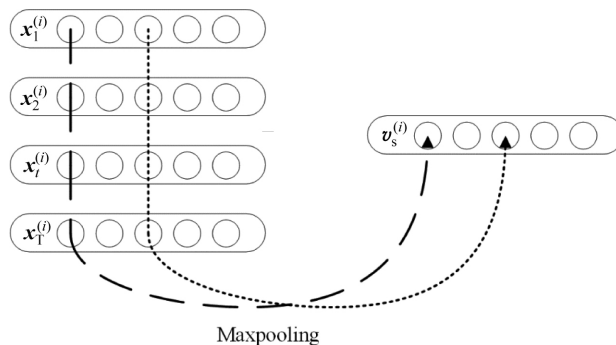


图2 使用 Maxpooling 提取目标向量

其中, $v_s^{(i)}$ 是句子 $s^{(i)}$ 的特定目标向量, $u_j^{(i)}$ 为 $v_s^{(i)}$ 的第 j 维, $x_{tj}^{(i)}$ 是句子 $s^{(i)}$ 中的第 t 个词项的短语向量的第 j 维的值。

同时,设置一个全局的目标向量 v_g 来表示“哪些单词对于分类目标更重要”,在训练过程中随机初始化并作为一个参数不断学习。

得到两个目标向量后,为了能够将单词的注解 $h_t^{(i)}$ 和目标向量相乘,需要通过一层全连接网络对 $h_t^{(i)}$ 进行处理,如式(6)所示。

$$h_t'^{(i)} = \tanh(W_s h_t^{(i)} + b_s) \quad (6)$$

然后对于句子中的所有单词,分别计算它和两个目标向量的相似度并归一化,得到针对两种目标向量的注意力得分,如式(7)~式(10)所示。

$$e_t^{(i)} = a(h_t'^{(i)}, v_g) = \exp(h_t'^{(i)T} v_g) \quad (7)$$

$$\alpha_t^{(i)} = \frac{e_t^{(i)}}{\sum_t e_t^{(i)}} \quad (8)$$

$$e_t'^{(i)} = a(h_t'^{(i)}, v_s^{(i)}) = \exp(h_t'^{(i)T} v_s^{(i)}) \quad (9)$$

$$\alpha_t'^{(i)} = \frac{e_t'^{(i)}}{\sum_t e_t'^{(i)}} \quad (10)$$

其中, $h_t'^{(i)}$ 是经过全连接网络处理后的单词表示, v_g 是训练得到的全局目标向量, $v_s^{(i)}$ 是在句子 $s^{(i)}$ 的词向量矩阵上利用最大池化得到的句子特有的目标向量, a 表示打分函数,这里具体使用的是点乘加指数函数, $h_t'^{(i)T}$ 表示 $h_t'^{(i)}$ 的转置。 $\alpha_t^{(i)}$ 和 $\alpha_t'^{(i)}$ 分别是单词注解 $h_t^{(i)}$ 对应 v_g 和 $v_s^{(i)}$ 两个目标向量的归一化后的分数。

将两个分数按照一定的权重相加作为最终的注意力分数,根据所有单词及其注意力分数得到句子的编码,如式(11)所示。

$$s^{(i)} = \sum_t ((1-\lambda)\alpha_t^{(i)} + \lambda\alpha'_t{}^{(i)})h_t^{(i)} \quad (11)$$

通过这样的方式,对于文档中的每个句子都能得到其对应的向量表示,并且文本里分类特征比较明显的词得到的权重会更大,在最终的句子表示中会占据主导地位。

2.3 文档编码及注意力层

在得到句子的向量表示 $s^{(i)}$ 后,类似句编码层,同样使用双向 GRU 对 $s^{(i)}$ 编码,如式(12)、式(13)所示。

$$\vec{h}^{(i)} = \overrightarrow{\text{GRU}}(s^{(i)}), i \in [1, L] \quad (12)$$

$$\overleftarrow{h}^{(i)} = \overleftarrow{\text{GRU}}(s^{(i)}), i \in [1, L] \quad (13)$$

然后连接 $\vec{h}^{(i)}$ 和 $\overleftarrow{h}^{(i)}$ 得到句子 $s^{(i)}$ 的注解 $h^{(i)}$, $h^{(i)}$ 包含了 $s^{(i)}$ 周围相邻句子的信息,但是仍然重点关注 $s^{(i)}$,如式(14)所示。

$$h^{(i)} = [\vec{h}^{(i)}, \overleftarrow{h}^{(i)}] \quad (14)$$

文档中的各个句子对文档分类结果的贡献程度也是不一样的,因此在文档注意力层,同样使用注意力机制来对每个句子打分。这里我们只关注文档的全局信息,因此使用一个文档级别的全局目标向量 v_d 来衡量句子的重要性。目标向量 v_d 表示了“哪些句子更重要”,在训练过程中同样被随机初始化并共同学习。

类似句注意力层,首先需要用全连接层对句子的注解 $h^{(i)}$ 进行处理,如式(15)所示。

$$h'^{(i)} = \tanh(W_d h^{(i)} + b_d) \quad (15)$$

然后根据文档级别的全局目标向量 v_d 计算相似度并归一化,得到注意力得分,然后通过加权的方式得到包含了文档中全部句子信息的文档向量 d ,如式(16)~式(18)所示。

$$e^{(i)} = a(h'^{(i)}, v_d) = \exp(h'^{(i)T} v_d) \quad (16)$$

$$\alpha^{(i)} = \frac{e^{(i)}}{\sum_i e^{(i)}} \quad (17)$$

$$d = \sum_i \alpha^{(i)} h^{(i)} \quad (18)$$

2.4 文档分类层

文档向量 d 是文档的高阶表示,可以直接用作文档分类的特征,通过 softmax 来计算每个类别的概率,如式(19)所示。

$$p = \text{softmax}(W_c d + b_c) \quad (19)$$

整个模型训练过程中使用交叉熵作为损失函数,如式(20)所示。

$$L = - \sum_d p_{d_j} \log p_{d_j} \quad (20)$$

其中, p_{d_j} 是文档 d 属于类别 j 的概率。

综上,HMAN 模型首先用双向 GRU 和混合的注意力机制对句子编码,得到文档中每个句子的向量表示;然后使用双向 GRU 和基础的注意力机制对文档编码。最终得到的文档表示包含了文档中所有句子的语义信息,而且重要句子占据了其中更大的比重。同样,在每个句子中,特征越明显的单词比重也越大。最后,在文档分类层,根据得到的文档表示进行分类,可以得到文档对应每个类别的概率。基于这样的层次结构可以得到类别特征更加明显的文档表示。

3 实验结果和分析

3.1 数据集及实验设置

为验证 HMAN 模型的效果,本文设计了两组实验,一组使用公开数据集,另一组是在企业年报数据这类具有层次结构的长文本数据集上进行验证,数据集的统计信息如表 1 所示。

表 1 实验数据集统计信息(单位:字)

数据集	文本	层次结构	平均长度	最大长度	最小长度
复旦大学	长	否	3 072	23 951	138
今日头条	短	否	32	161	6
企业年报	长	是	3 451	33 062	161

第一组实验采用的是复旦大学中文文本分类数据集和 GitHub 官网下载的今日头条短文本分类数据集。复旦大学中文文本分类数据集共有 20 个类别,9 832 条数据。今日头条短文本分类数据集共有 382 688 条数据,15 个类别,将“新闻标题”信息作为文本数据,“分类名称”信息作为分类标签。

第二组实验采用的上市公司年报数据集共有 31 230 条企业年报数据,以“董事会讨论”信息作为文本数据,分别提取一级分类和二级分类作为分类标签进行文本分类,其中一级分类共有 18 类,二级分类共有 78 类。

在训练过程中,将数据集的全部数据按 9:1 划分为训练集和测试集,训练集中又取 10% 作为开发集。

实验将 HMAN 模型与以下分类模型进行比较:

(1) 基于机器学习的分类模型: 贝叶斯模型、决策树模型。其中贝叶斯模型分别训练伯努利贝叶斯分类器^[1]和多项式朴素贝叶斯分类器^[16], 决策树模型分别训练基于信息熵的 ID3 决策树^[17]和基于 GINI 不纯度的 CART 决策树^[18]。

(2) 基于深度学习的分类模型: 包括不使用层次模型的 TextCNN^[5]、TextRNN^[19], 以及使用层次模型的 HAN^[8]。

HMAN 模型的部分训练参数设置如表 2 所示。其中批量样本的大小设为 32, GRU 中隐藏单元的个数为 50, 词向量的维度为 200, Adam 优化算法中的学习率为 0.001, 梯度裁剪的阈值为 5。句注意力层的超参数 λ 经过调整和验证, 在 $\lambda=0.2$ 时取得的效果最好。

表 2 HMAN 部分训练参数设置

参数类型	参数值
epoch	100
batch_size	32
hidden_size	50
embedding_size	200
learning_rate	0.001
grad_clip	5

3.2 公开数据集上的实验结果与分析

第一组实验对比了传统机器学习算法、卷积神经网络文本分类模型 TextCNN、循环神经网络文本分类模型 HAN 以及本文提出的 HMAN, 实验结果如表 3 所示, 其中, TextCNN 模型、HAN 模型与

表 3 各模型在公开数据集上的分类结果

模型	准确率	
	复旦大学公开数据集	今日头条公开数据集
多项式朴素贝叶斯	0.743 90	0.875 67
伯努利贝叶斯	0.677 85	0.847 71
KNN($K=1$)	0.873 98	0.631 35
GINI 不纯度决策树	0.880 08	0.828 87
信息熵决策树	0.876 02	0.789 36
TextCNN	0.932 92	0.876 08
HAN	0.940 39	0.876 89
HMAN	0.954 27	0.878 05

HMAN 模型取分类效果收敛之后的十次平均分类准确率作为分类结果。

朴素贝叶斯模型假设属性之间是相互独立的, 但是在实际应用中, 这个假设往往不成立, 在属性数目比较多时, 朴素贝叶斯的性能表现不好。KNN 算法在 K 取值为 1 时, 在复旦大学中文数据集上取得的效果较好, 在今日头条短文本分类数据集上的效果较差, KNN 算法的 K 值与数据集本身有很大的关系, K 值的确定目前还没有特定的经验公式, 因此在对新的数据集进行训练时, 需要重新对 K 取不同的值进行实验来确定最佳 K 值。在两组实验数据集上, GINI 不纯度决策树的分类准确率比信息熵决策树的分类准确率要高。这是由于信息熵对不纯度的敏感度要高于基尼系数, 因此当把信息熵作为指标时, 决策树的生长会更加“精细”, 在数据的维度高或者数据噪声大时, 信息熵很容易过拟合。

TextCNN 在两组数据集上都取得了很好的分类效果, 均高于传统机器学习分类模型的准确率。HAN 分类效果略高于 TextCNN。HMAN 模型在公开数据集上取得了与 HAN 相当的结果, 在复旦大学数据集上, HMAN 模型的分类准确率达到 95.43%, 略好于 HAN 模型, 差异值为 0.014 左右。在今日头条短文本分类数据集上, HMAN 模型的分类准确率达到 87.81%, 和 HAN 模型效果相当, 差异值为 0.001 左右。相比短文本, HMAN 模型更能关注到长文本中最具区别性的语义特征, 效果更好。

3.3 上市公司年报数据集上的实验结果与分析

为验证 HMAN 模型在具有层次结构的长文本数据集上的性能, 本节采用上市公司的年报数据库数据, 企业年报数据库提供了企业董事会信息和经营产品信息包括: 记录 ID、股票代码、股票简称、行业分类、产品名称、董事会讨论、主要产品、经营范围、年份等字段。其中, “董事会讨论”是企业董事会讨论与分析的文本内容, “行业分类”根据不同粒度划分为四个级别。

本组实验以“董事会讨论”信息作为文本数据, 分别提取一级分类和二级分类作为分类标签进行文本分类, 确定企业所属类别。其中一级分类共有 18 类, 二级分类共有 78 类。企业年报数据示例如表 4 所示。

表 4 企业年报数据示例

序号	董事会讨论	一级分类	二级分类
R104-2003	公司属光纤光缆制造行业。2003 年,全球光缆产业持续下滑以及光缆市场竞争的进一步加剧,使整个行业仍未有大的起色。全国 500 多家光缆企业萎缩至 50 来家,大部分光缆企业处于停产状态……	制造业	电气机械及器材制造业
R108-2012	管理层讨论与分析:公司报告期内,由于受经济不景气的影响,且“宽带中国战略”的实施方案推迟出台,导致通信运营商投资需求增长有限,招标计划延后,订单交付推迟;同时,受公司经营成本上升的影响,公司经营业绩下滑……	制造业	计算机、通信和其他电子设备制造业
R005-2013	一、董事会关于公司报告期内经营情况的讨论与分析:2013 年,在公司管理层带领下,百视通继续做大“四屏”(电视、PC、手机、Pad)新媒体主营业务,扩大公司营收规模,同时公司大力拓展智能电视与互联网产业链……	信息传输、软件和信息技术服务业	电信、广播电视和卫星传输服务

各模型在一级行业分类和二级行业分类的准确率如表 5 所示。可以看出,基于深度学习的文本分类模型整体效果好于基于机器学习的分类模型。一方面,基于机器学习的分类模型会依赖某种规则,约束了分类的准确率;另一方面,特征工程会对基于机器学习的分类模型的分类结果产生很大影响,这里使用了基于文档的统计信息的词频和逆文档频率作为文本特征,不能表示文本的语义。而基于深度学习的文本分类模型使用了包含单词上下文信息的语义特征,同时用 CNN 或 RNN 能进一步提取文本序列的特征,更有助于对文本进行分类。

表 5 各模型行业分类准确率(%)

分类模型	一级分类	二级分类
多项式朴素贝叶斯	56.9	26.6
伯努利贝叶斯	68.6	33.5
GINI 不纯度决策树	70.2	49.8
信息熵决策树	66.4	36.6
TextCNN	75.7	57.0
TextRNN	75.8	59.2
HAN	83.9	72.4
HMAN	87.3	74.8

对于不使用层次模型的 TextCNN 和 TextRNN,TextCNN 使用不同大小的卷积核可以提取文档中不同窗口大小的固定特征,而 TextRNN 可以捕捉文本序列中的依赖信息,同时注意力机制的使用可以使编码后的文档向量更关注文本中的重要特征。两者均有各自的优势,在一级分类的结果上不分伯仲,而在类别较多的二级分类上 TextRNN

略好于 TextCNN。另一方面,在训练效率上 TextCNN 相对 TextRNN 要快很多。

使用层次模型可以进一步提高分类效果。HAN 和 HMAN 在一级分类和二级分类上的准确率均明显高于不使用层次模型的 TextCNN、TextRNN。这说明引入文档的层次结构可以在不同层级分别关注文本的句子特征和文本特征,有助于提升文档级别的文本分类的效果。

本文提出的 HMAN 在句子级别的注意力机制中,在全局目标向量的基础上,从词向量矩阵中利用最大池化提取句子中每个维度上的最大特征作为目标向量,能够更加突出具有明显类别特征的语义信息,分类效果好于 HAN。

另外,为了观察基于深度学习的分类模型在训练过程中的变化趋势,以一级分类为示例绘制了各模型在开发集上的准确率变化曲线,如图 3 所示。从图 3 中可以看出,由于直接通过提取的方式获得目标向量,不需要额外的参数训练,HMAN 的收敛速度比 HAN 快。

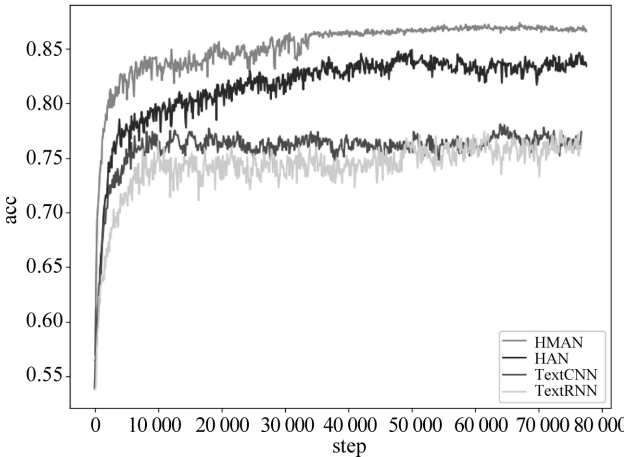


图 3 一级分类准确率变化曲线

从表 5 也可以看出,无论采用何种分类模型,二级分类的准确率都明显低于一级分类准确率。主要原因在于:一级分类只有 18 类,而二级分类有 78 类,类别数量越多,分类产生的错误可能就越多。由于一级分类粒度较大,各类别的区分度比较明显,因此比较容易捕获文本的类别特征。由于二级分类是在一级分类的基础上进行的细分,同一大类的各个小类的区分性也不大,特征不够明显,因此会对分类的准确度造成影响。

综合上述分析,HMAN 模型在分类准确率和模型的性能上都达到了预期的实验效果,在实验数据集为结构化长文本时,更能体现出 HMAN 模型的分类型优势。

4 结论

本文研究基于深度学习的文本分类算法,提出基于层次混合注意力机制的文本分类模型 HMAN。首先,改进分层注意力模型,在句编码时,提出为文本中每个句子设置特定的目标向量,结合全局目标向量,按照一定的权重对所有词项及其注意力得分进行综合得到句编码。然后,通过分层注意力模型中文档编码层、文档注意力层和文档分类层,获得文档对应每个类别的概率,即实现文本分类。通过对公开数据集和企业年报数据上的数据进行行业分类,实验表明 HMAN 在一级分类和二级分类上的分类准确率均好于传统机器学习的文本分类模型和当前已有的基于深度学习的文本分类模型。

参考文献

- [1] Lewis D D. Naive (Bayes) at forty: The independence assumption in information retrieval [M]. Berlin Heidelberg: Springer, 1998: 4-15.
- [2] Cover T, Hart P. Nearest neighbor pattern classification [J]. IEEE Transactions on Information Theory, 2003, 13(1): 21-27.
- [3] Cortes C, Vapnik V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273-297.
- [4] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述[J]. 计算机学报, 2017(6): 1229-1251.
- [5] Yoon Kim. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014: 1746-1751.

- [6] Liu Pengfei, Qiu Xipeng, Huang Xuanjing. Recurrent neural network for text classification with multi-task learning[C]//Proceedings of the International Joint Conferences on Artificial Intelligence, 2016: 2873-2879.
- [7] Tang Duyu, Qin Bing, Liu Ting. 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.
- [8] Yang Zichao, Yang Diyi, Dyer Chris. 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.
- [9] 程艳, 叶子铭, 王明文等. 融合卷积神经网络与层次化注意力网络的中文文本情感倾向性分析[J]. 中文信息学报, 2019, 33(1): 133-142.
- [10] Xiao Y, Cho K. Efficient Character-level document classification by combining convolution and recurrent layers [J/OL]. arXiv preprint arXiv: 1602.00367, 2016.
- [11] Conneau A, Schwenk H, Barrault, Loïc, et al. Very deep convolutional networks for text classification [C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2016: 1107-1116.
- [12] Rie Johnson, Tong Zhang. Deep pyramid convolutional neural networks for text categorization[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 562-570.
- [13] Xu Jiacheng, Chen Danlu, Qiu Xipeng, et al. Cached long short term memory neural networks for document-level sentiment classification[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016: 1660-1669.
- [14] Mnih V, Heess N, Graves A, et al. Recurrent models of visual attention[J]. Advances in Neural Information Processing Systems, 2014(3): 2204-2212.
- [15] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [C]//Proceedings of the 2015 International Conference on Learning Representations, 2015: 1-15.
- [16] Quinlan J R. Learning efficient classification procedures and their application to chess end games[M]. Machine Learning: An Artificial Intelligence Approach, 1983(1): 463-482.
- [17] McCallum A, Nigam K. A comparison of event models for naive Bayes text classification[G]//Proceedings of AAAI-98 Workshop on Learning for Text

Categorization, 1998, 752(1): 41-48.

- [18] Breiman L, Friedman J, Stone J C, et al. Classification and regression trees [M]. London: Chapman and Hall/CRC, 1984.



孙新(1975—), 博士, 副教授, 主要研究领域为人工智能、机器学习。

E-mail: sunxin@bit.edu.cn



赵永妍(1997—), 硕士研究生, 主要研究领域为文本分类。

E-mail: 3220190927@bit.edu.cn

- [19] Zhang X, Zhao J, LeCun Y. Character level convolutional networks for text classification[C]//Proceedings of the 2015 Conference and Workshop on Neural Information Processing Systems, 2015: 649-657.



唐正(1994—), 硕士研究生, 主要研究领域为机器学习、智能问答。

E-mail: 2120171060@bit.edu.cn

(上接第 68 页)

- [20] Zhang J, Liu S, Li M, et al. Bilingually-constrained phrase embeddings for machine translation[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, 1: 111-

121.

- [21] 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.



颜欣(1993—), 硕士研究生, 主要研究领域为复述抽取与生成。

E-mail: xyan@ir.hit.edu.cn



潘晓彤(1984—), 学士, 高级工程师, 主要研究领域为自然语言处理、对话系统。

E-mail: panxiaotong@xiaomi.com



张宇(1972—), 通信作者, 博士, 教授, 主要研究领域为自然语言处理、问答系统、个性化信息检索。

E-mail: zhangyu@ir.hit.edu.cn