



## 融合知识图谱与注意力机制的短文本分类模型

丁辰晖<sup>1</sup>, 夏鸿斌<sup>1,2</sup>, 刘 渊<sup>1,2</sup>

(1. 江南大学 数字媒体学院, 江苏 无锡 214122; 2. 江苏省媒体设计与软件技术重点实验室, 江苏 无锡 214122)

**摘要:** 针对短文本缺乏上下文信息导致的语义模糊问题, 构建一种融合知识图谱和注意力机制的神经网络模型。借助现有知识库获取短文本相关的概念集, 以获得短文本相关先验知识, 弥补短文本缺乏上下文信息的不足。将字符向量、词向量以及短文本的概念集作为模型的输入, 运用编码器-解码器模型对短文本与概念集进行编码, 利用注意力机制计算每个概念权重值, 减小无关噪声概念对短文本分类的影响, 在此基础上通过双向门控循环单元编码短文本输入序列, 获取短文本分类特征, 从而更准确地进行短文本分类。实验结果表明, 该模型在 AGNews、Ohsumed 和 TagMyNews 短文本数据集上的准确率分别达到 73.95%、40.69% 和 63.10%, 具有较好的分类能力。

**关键词:** 短文本分类; 知识图谱; 自然语言处理; 注意力机制; 双向门控循环单元

开放科学(资源服务)标志码(OSID):



**中文引用格式:** 丁辰晖, 夏鸿斌, 刘渊. 融合知识图谱与注意力机制的短文本分类模型[J]. 计算机工程, 2021, 47(1): 94-100.

**英文引用格式:** DING Chenhui, XIA Hongbin, LIU Yuan. Short text classification model combining knowledge graph and attention mechanism[J]. Computer Engineering, 2021, 47(1): 94-100.

### Short Text Classification Model Combining Knowledge Graph and Attention Mechanism

DING Chenhui<sup>1</sup>, XIA Hongbin<sup>1,2</sup>, LIU Yuan<sup>1,2</sup>

(1. School of Digital Media, Jiangnan University, Wuxi, Jiangsu 214122, China; 2. Jiangsu Key Laboratory of Media Design and Software Technology, Wuxi, Jiangsu 214122, China)

**[Abstract]** Concerning the semantic ambiguity caused by the lack of context information, this paper proposes a neural network model, which combines knowledge graph and attention mechanism. By using the existing knowledge base to obtain the concept set related to the short text, the prior knowledge related to the short text is obtained to address the lack of context information in the short text. The character vector, word vector, and concept set of the short text are taken as the input of the model. Then the encoder-decoder model is used to encode the short text and concept set, and the attention mechanism is used to calculate the weight value of each concept to reduce the influence of unrelated noise concepts on short text classification. On this basis, a Bi-directional-Gated Recurrent Unit (Bi-GRU) is used to encode the input sequences of the short text to obtain short text classification features, so as to perform short text classification more effectively. Experimental results show that the accuracy of the model on AGNews, Ohsumed and TagMyNews short text data sets is 73.95%, 40.69% and 63.10%, respectively, showing a good classification ability.

**[Key words]** short text classification; knowledge graph; Natural Language Processing (NLP); attention mechanism; Bi-directional-Gated Recurrent Unit (Bi-GRU)

**DOI:** 10. 19678/j. issn. 1000-3428.0056734

### 0 概述

近年来,随着 Twitter、微博等社交网络的出现,人们可以轻松便捷地在社交平台上发布文本、图片、视频等多样化的信息,社交网络已超越传统媒体成为新的信息聚集地,并以极快的速度影响着社会的信息传播格局<sup>[1]</sup>。如何对这些短文本进行准确分类,是自然语

言处理(Natural Language Processing, NLP)领域中的一项关键技术。由于这些短文本篇幅较短,缺乏上下文信息,且内容口语化、特征属性多与噪声较大,因此精确提取文本特征,采用合适的分类模型对短文本进行分类是一个亟需解决的问题。

在文本分类这一领域中,一般的文本表示方法

**基金项目:** 国家自然科学基金(61672264); 国家科技支撑计划项目(2015BAH54F01)。

**作者简介:** 丁辰晖(1994—),男,硕士研究生,主研方向为自然语言处理;夏鸿斌,副教授、博士;刘 渊,教授、博士生导师。

**收稿日期:** 2019-11-28 **修回日期:** 2020-01-15 **E-mail:** 1627617238@qq.com

分为显式表示与隐式表示。对于显式表示方法,人们一般从知识库、词性标注、句法分析<sup>[2]</sup>等多个方面创造有效的特征,将短文本表示为稀疏向量,每一维度都是显式的特征。虽然文本的显式表示很容易被人理解,但是显式表示往往忽略了短文本的上下文,无法捕捉到深层的语义信息,此外还存在数据稀疏问题。例如,当实体特征在知识库中不存在时,则无法获得它的任何特征,此时显式表示将无法工作。现阶段在深度学习中隐式文本的隐式表示方法则更为常见,通过训练词向量将每个词映射成为密集的向量<sup>[3]</sup>,使用词向量矩阵表示短文本,由于词向量中包含词义信息,神经网络模型可以从上下文中获取更丰富的语义信息,促进神经网络模型对短文本的理解。但是隐式表示方法仍然存在一些缺点,如短文本为{The Bulls won the NBA championship},在文中Bulls是一个篮球队的名字,然而通过词向量输入的模型可能无法捕捉到这一信息,将其视为一种动物或一个新词,造成分类效果不够理想。单纯使用显示或隐式的文本表示方法都存在的问题,所以将两者相结合,利用一个内容丰富的知识库来丰富短文本的先验知识,获取短文本的概念集,再将短文本与概念集映射为词向量矩阵,从而使模型学习出更全面、更深层的语义,提升分类能力。

本文构建一种融合知识图谱和注意力机制的神经网络模型。将知识图谱与短文本分类模型相融合,从已有知识库中获取短文本的概念集作为输入,从而获得文本中的先验知识。在此基础上引入注意力机制,计算每个概念相对于概念集及短文本之间的相关性,对两者注意力权重进行加权融合,得出最终每个概念的权重,以提高相关概念的权重,使模型分类效果更具判别性。

## 1 相关工作

随着深度学习的发展,越来越多的学者使用深度学习方法进行文本分类的研究,其在绝大多数任务中的表现都优于传统方法,极大地促进了文本分类这一领域的发展。因为卷积神经网络(CNN)在NLP和计算机视觉的各个领域都表现出了较好的性能,所以受到了研究人员的极大关注。文献[4]利用预训练的词向量,将卷积神经网络应用在语句分类任务中。文献[5]提出CNN动态的k-max pooling方法来解决Twitter短文本的极性分类问题,CNN可以处理不同长度的输入句子,并在句子上生成一个特征图,能够明确捕捉短期和长期关系,并取得了较好的效果。文献[6]使用字符级卷积神经网络进行文本分类,利用字符作为模型输入从而代替了词语作为输入,并在实验数据集上取得了良好的效果。文献[7]采用循环神经网络建立篇章级循环神经网络模型,该模型相比标准的循环神经网络模型具有更

强的性能,在文本分类任务中具有较好的效果。

2017年,Google团队<sup>[8]</sup>提出使用注意力机制的Transformer模型解决NLP问题,随着注意力机制在NLP领域中的广泛应用,越来越多的学者开始利用注意力机制解决NLP方面的问题。2018年,HUANG等人<sup>[9]</sup>构建AOA\_LSTM模型,该模型使用双向LSTM构建了句子的属性特征向量矩阵,并通过上下文编码和注意力计算更好地关注属性序列中的重要信息,挖掘出更深层的情感特征信息。2018年,文献[10]使用双向门控循环单元(Bi-directional-Gated Recurrent Unit, Bi-GRU)结合注意力机制,在餐饮电商评论短文本分类中取得了较好的效果。这些方法证明了深度学习结合注意力机制在短文本情感分类中可以取得更好的效果。

2016年,微软研究院发布了概念图谱。概念图谱是一个大型的知识图谱系统,通过对来自数以亿计的网页和数年积累的搜索日记的数据进行学习而掌握大量的常识性知识。概念图谱表示形式为实例、概念和关系的三元组。实例与概念之间为IsA关系,如三元组(苹果,水果,IsA)表示苹果是一种水果。文献[11]提出了6种基于概念图谱进行实例概念化的方法,并在微软知识图谱的官方网站提供了相应api函数的调用。

2017年,WANG等人<sup>[12]</sup>提出一种融合概念图谱的CNN短文本分类模型,通过从知识库中提前获取的短文本先验知识与CNN提取的文本特征相结合,可以一定程度上解决短文本分类中缺乏上下文信息的缺点,在5个公开数据集中获得了优异的效果。由此可见,通过知识图谱与深度学习相结合,可以缓解短文本缺乏上下文的问题,融合知识图谱的模型可以获取词向量以外的额外信息,在文本分类任务中具有较好的表现。

虽然融合了知识图谱的神经网络模型有着较好的表现,但是仍然存在一些问题。例如上文中{The Bulls won the NBA championship},虽然可以从知识库中获取到球队和动物这两个概念,但显然动物在短文本中是不恰当的概念,这些噪声会影响文本分类的结果。此外,例如输入短文本为{Steve Jobs is one of the co-founders of Apple},可以从知识库中检索到乔布斯的企业家和个人两个概念,虽然两个概念都正确,但显然在短文本中企业家的概念应该占更大的权重。

为解决上述问题,本文引入了注意力机制,同时借鉴Transformer模型,提出一种融合知识图谱的注意力门控循环单元网络,通过计算短文本与其概念集中的概念的注意力权重,赋予与短文本密切相关的概念更高的权重,如上例中{Steve Jobs is one of the co-founders of Apple},Steve Jobs的企业家和个人两个概念,增大企业家概念的权重,减小个人概念

的权重,使得文本分类模型更具有判别性。

## 2 知识增强的文本分类模型

本文提出了一种融合知识图谱、注意力机制和双向GRU的知识增强网络模型(Knowledge Enhanced Attention Bi-GRU, KEAT-GRU),该模型借鉴了神经网络翻译模型的设计思想<sup>[8]</sup>,采用基于Transformer的编码器-解码器网络结构,同时融合概念图谱获取短文本的先验知识,如图1所示,该网络模型主要由

以下两部分构成:

1)短文本编码:使用字符向量与词向量拼接后作为输入,经过Bi-GRU提取短文本特征,并利用多头自注意力层对重要文本信息进行加权,获得短文本特征。

2)概念化编码:通过调用微软概念图谱的API,获取短文本概念集并向量化,通过与短文本的特征向量进行Attention计算,提升概念集中与短文本关系密切的概念的权重,最终得出概念集特征。

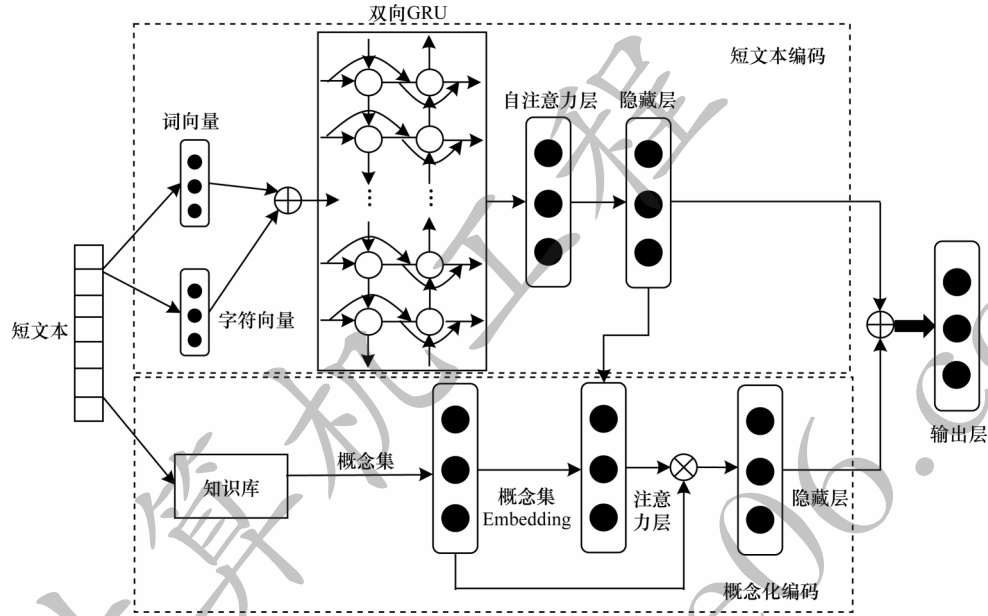


图1 文本分类模型结构

Fig.1 Structure of text classification model

### 2.1 短文本编码

已有研究表明<sup>[13]</sup>,使用卷积神经网络可以提取出单词字符中的形态学信息(例如单词的前缀后缀),将字符嵌入作为词向量的拓展,为缺少词向量的单词提供额外信息。因此,本文使用字符级词嵌入向量与词向量相拼接作为短文本编码模型的输入,输入短文本单词序列 $\{x_1, x_2, \dots, x_n\}$ , $x_i$ 表示句中第 $i$ 个单词,其中, $x_i$ 单词中包含长度为 $L$ 的字符, $c_j$ 为单词 $x_i$ 中每个字符嵌入向量,每一个字符都代表其相应的一个特征。如图2所示,使用一个标准卷积神经网络处理每一个单词中的字符序列,训练得出单词的字符级向量 $e_i^{w_c}$ ,计算公式如式(1)所示:

$$e_i^{w_c} = \max_{1 \leq j \leq L} \left( W_{CNN}^T \begin{bmatrix} e^c \left( c_{j - \frac{ke-1}{2}} \right) \\ \vdots \\ e^c \left( c_{j - \frac{ke-n}{2}} \right) \end{bmatrix} + b_{CNN} \right) \quad (1)$$

其中, $W_{CNN}$ 与 $b_{CNN}$ 为训练参数,ke表示卷积核大小,max表示进行最大池化操作。

随后模型将单词 $x_i$ 映射为词向量 $e^w$ :

$$e_i^w = E(x_i) \quad (2)$$

对词向量与字符向量进行拼接:

$$E_i = \text{Concat}(e_i^w; e_i^{w_c}) \quad (3)$$

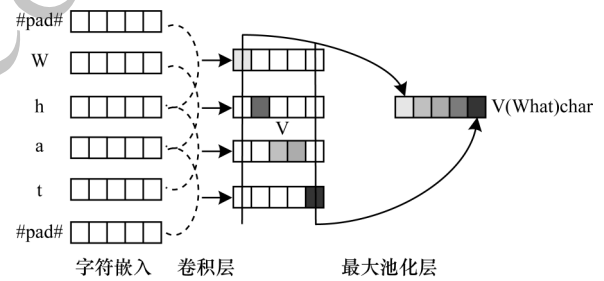


图2 字符级卷积神经网络结构

Fig.2 Structure of character-level convolution neural network

最终获得词向量矩阵 $E=[E_1, E_2, \dots, E_n]$ 作为Bi-GRU的输入。前向GRU按照正常的顺序读取输入序列( $E_1 \sim E_n$ ),反向GRU则按逆序读取输入序列( $E_n \sim E_1$ ),每个 $t$ 时刻的输入向量 $E_t$ 经过门控循环单元的计算,获取每个时刻的前向隐藏状态( $\vec{h}_1, \vec{h}_2, \dots, \vec{h}_t$ )和反向隐藏状态( $\tilde{h}_1, \tilde{h}_2, \dots, \tilde{h}_t$ ),将各个时刻的前向隐藏状态 $\vec{h}_j$ 与其对应时刻的反向隐藏状态 $\tilde{h}_j$ 连接,得到该时



刻的隐藏状态:

$$\mathbf{h}_j = [\vec{\mathbf{h}}_j \parallel \overleftarrow{\mathbf{h}}_j]^T \quad (4)$$

随后将每个时刻隐藏状态  $\mathbf{h}_j$  输入自注意力层,对每个时间步输入的词根据注意力计算进行加权,使重要的词语获得更高的权重,Attention 计算定义为:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (5)$$

其中,  $\mathbf{Q}$  表示一次执行 Attention 时的查询,  $\mathbf{K}$  表示与值相对应同时又用来与查询计算相似度作为 Attention 选取的依据,  $\mathbf{V}$  表示被注意并被选取的数据。  $\mathbf{Q} \in \mathbb{R}^{n \times d_k}$ ,  $\mathbf{K} \in \mathbb{R}^{m \times d_k}$ ,  $\mathbf{V} \in \mathbb{R}^{m \times d_v}$ , 输入包含  $d_k$  维的 query 和 key 以及  $d_v$  维的 value。

Multi-head Attention 运算结构如图 3 所示。

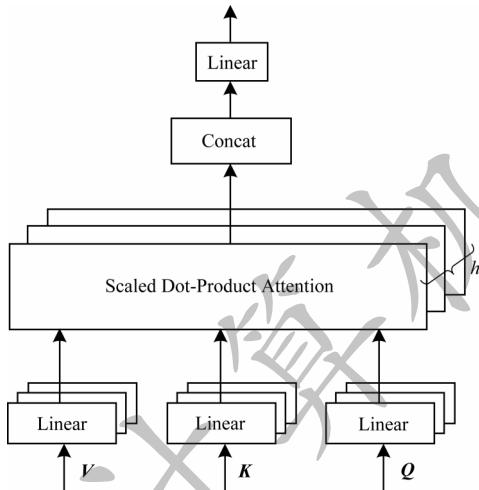


图3 Multi-head Attention 运算结构

Fig.3 Operation structure of Multi-head Attention

计算定义如下:

$$\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) \quad (6)$$

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h) \quad (7)$$

$$\mathbf{a}_i = \text{MultiHead}(\mathbf{h}_i, \mathbf{h}_i, \mathbf{h}_i) \quad (8)$$

用  $h$  个不同的线性变换分别将  $d_{\text{model}}$  维的 key、value 和 query 映射成  $d_k$  维、 $d_k$  维和  $d_v$  维,随后计算得出  $h \times d_v$  维输出,进行拼接,最后进行一次线性变换得到最终的输出。 $\mathbf{h}_i$  为输入序列,即 Bi-GRU 层输出的隐藏状态,目的是在输入序列内部进行注意力计算,寻找序列内部的联系。

通过自注意力层计算出注意力权重  $\mathbf{a}_i$ ,将 Bi-GRU 输出的  $t$  时刻的隐藏状态  $\mathbf{h}_i$  加权平均:

$$\mathbf{h}'_i = \sum_{t=1}^T \mathbf{a}_t \mathbf{h}_t \quad (9)$$

最终输出特征矩阵  $\mathbf{h}' \in \mathbb{R}^{n \times 2u}$ 。

## 2.2 短文本概念化编码

文本概念化需要通过已有知识库例如 Yago<sup>[14]</sup>、Microsoft Concept Graph。本文使用微软发布的 Concept Graph 知识图谱对短文本进行概念化,获取

文本相关概念集。将每条短文本通过知识库获取文本的概念集合  $\mathbf{C} = (c_1, c_2, \dots, c_m)$ ,  $c_i$  表示第  $i$  个概念集中的概念向量。为增加重点概念向量的权重,减小与短文本无关的概念向量对结果的影响,首先将短文本特征矩阵  $\mathbf{h}' \in \mathbb{R}^{n \times 2u}$  经过最大池化层,转换为特征向量  $\mathbf{q} \in \mathbb{R}^{2u}$ ,随后引入注意力机制,计算概念集中第  $i$  向量与其短文本特征向量  $\mathbf{q}$  的关系权重:

$$\alpha_i = \text{softmax}(\mathbf{v}_i^T \tanh(\mathbf{W}_1 \times \text{concat}[\mathbf{c}_i; \mathbf{q}] + b_1)) \quad (10)$$

其中,  $\alpha_i$  为第  $i$  个概念集中的概念向量与其短文本之间的注意力权重,  $\mathbf{W}_1 \in \mathbb{R}^{d_a \times (2u+d)}$  为权重矩阵,  $\mathbf{v}_i \in \mathbb{R}^{d_a}$  为权重向量,  $d_a$  为超参数,  $b_1$  为偏置。

在概念集内部加入自注意力机制并进行注意力计算,以获取每个概念  $c_i$  在整个概念集中的重要性权重:

$$\beta_i = \text{softmax}(\mathbf{v}_i^T \tanh(\mathbf{W}_2 \mathbf{c}_i) + b_2) \quad (11)$$

其中,  $\beta_i$  为第  $i$  个概念集中的概念向量的注意力权重,  $\mathbf{W}_2 \in \mathbb{R}^{d_b \times d}$  为权重矩阵,  $\mathbf{v}_i \in \mathbb{R}^{d_b}$  为权重向量,  $d_b$  为超参数,  $b_2$  为偏置,注意力机制赋予重要概念较大的权重,赋予不重要的概念极小的权重(接近于零),以突出概念集中重要概念。

在得到  $\alpha_i, \beta_i$  注意力权重后,用式(12)将两者结合,以获取最终注意力权重:

$$\mathbf{a}_i = \text{softmax}(\gamma \alpha_i + (1-\gamma) \beta_i) \quad (12)$$

其中,  $\mathbf{a}_i$  为最终第  $i$  个概念向量注意力权重,  $\gamma \in [0, 1]$  是调节  $\alpha_i$  与  $\beta_i$  的权重参数。

获取每个概念向量的注意力权重后,对每个概念向量进行加权计算:

$$\mathbf{r} = \sum_{i=1}^m \alpha_i \mathbf{c}_i \quad (13)$$

## 2.3 模型训练

本文网络模型训练采用反向传播算法,同时引入 L2 正则化以避免网络模型过拟合问题。L2 正则化通过在损失函数中加入 L2 范数作为惩罚项,使得模型拟合更倾向于低维的模型,可以有效防止过拟合。相比于 L1 正则化会产生稀疏性问题, L2 正则化可以使系数向量更加平滑,避免稀疏性问题。本文通过最小化交叉熵损失函数来优化网络模型,完成分类任务,交叉熵损失函数为:

$$\text{loss} = - \sum_{i=1}^D \sum_{j=1}^C y_i^j \ln y_i^j + \lambda \|\theta\|^2 \quad (14)$$

其中,  $D$  为训练集大小,  $C$  为类别数,  $y$  为预测类别,  $y'$  为实际类别,  $\lambda \|\theta\|^2$  为正则项。

## 2.4 实验数据集

将本文方法在两个不同领域数据的公开数据集上进行实验,解决短文本情感分析的任务。

1) SemEval2017 数据集是国际语义评测比赛 Task4 的 Twitter 数据集,数据集的短文本中共包含积极、中性和消极 3 种情感分类。

2) AGNews: 文献[15]通过互联网获得了新闻

文章语料库,其中包含来自2 000多个新闻来源的496 835个分类新闻文章,仅从标题和描述字段中选择该语料库中最大的4个类来构建数据集。

3) Ohsumed: 文献[16]发布的医学文献书目分类数据集,删除了带有多个标签的文档,仅将标题用于短文本分类。

4) 国际计算机语言协会(Association for Computational Linguistics, ACL)公布的电影评论数据集和IMDB影评数据集,每一条评论包含正向和负向的感情倾向。

5) TagMyNews: 文献[17]发布的英文新闻文本数据集,使用其新闻标题作为数据集,包含政治、艺术等7个分类。

语料库统计如表1所示。

表1 语料库统计  
Table 1 Corpora statistics

数据集	文档数	分类数	平均句长
Twitter	49 570	3	12.3
AGNews	6 000	4	18.4
Ohsumed	7 400	23	6.8
Movie Review	9 584	2	8.6
TagMyNews	32 549	7	5.1

## 2.5 对比方法

各数据集对比方法如下:

1) CNN。文献[4]提出的卷积神经网络模型,是较为基础的卷积神经网络。

2) AT-CNN(Attention-based CNN)。文献[18]提出基于词注意力卷积神经网络,在词嵌入层后加入注意力层,可以获取较为重要的局部特征,取得较好的结果。

3) CNNs-LSTMs。文献[19]提出的CNN与LSTM相结合的模型,使用未标注数据集训练词向量,运用标注数据集在训练中微调参数,最终在Twitter情感分类任务中取得优秀的成绩。

4) AT-LSTM(Attention-based LSTM)。文献[20]提出基于注意力机制的双向LSTM网络,该模型在Twitter的情感分类中取得了比传统LSTM网络更好的分类效果。

5) KCNN(Knowledge Convolutional Neural Network)。文献[9]提出融合概念图谱的CNN短文本分类模型,将知识库中提前获取的短文本先验知识与CNN提取的文本特征相结合,该模型在5个公开数据集上获得优异的效果。

## 2.6 实验参数设置

模型使用的预训练词向量维度为300,字符向量维度为50维,概念向量维度为100维,权重随机初始化标准差为0.1的正态分布随机数。同时所有词向量、字符向量及概念向量都在训练时进行微调。在词嵌入层,池化层设置dropout值为0.3。隐藏层维度为100,L2正则化权重为0.000 1,学习率为0.01,

学习率的下降率为0.05。使用Adam优化方法加快模型训练速度,通过每个batch中50个样本的方式进行模型的训练。

## 2.7 实验结果与分析

将本文模型与5种对比模型在5个公开数据集上进行实验,结果如表2所示。

表2 各数据集准确率实验结果

Table 2 Experimental results of accuracy of each data set

模型	Twitter	AGNews	Ohsumed	Movie Review	TagMyNews
CNN	0.614 5	0.672 4	0.329 2	0.756 6	0.571 2
AT-CNN	0.648 8	0.683 2	0.342 4	0.773 8	0.594 3
CNNs-LSTMs	0.653 6	0.679 8	0.336 0	0.768 4	0.574 5
AT-LSTM	0.659 3	0.684 5	0.347 8	0.771 5	0.584 7
KCNN	0.661 9	0.713 6	0.389 7	0.779 7	0.624 1
KEAT-GRU	0.670 2	0.739 5	0.406 9	0.784 7	0.631 0

从表2可以看出,本文提出的方法模型在不同领域的数据集上都取得了较好的效果,尤其在Ohsumed、TagMyNews数据集中表现优异,这是由于这些数据集只使用新闻、文章标题构建,文本长度过短,缺乏上下文信息,在没有先验知识的情况下模型训练效果较差,由于在出现特有名词或者新词时,可能这些词在预训练词向量集中不存在,因此模型一般只能使用随机初始化的方法处理这一问题,这会导致模型分类判别效果变差。而KCNN与本文模型由于都融入了知识图谱,可以获取文本中的先验知识,因此都取得了优异的分类效果。但是本文模型在Twitter与Movie Review数据集上提升不是很明显,主要有两方面原因,一方面是因为这两个数据集是用户的博文和评论,可能含有一定的上下文信息,另一方面因为两个数据集是情感分类方向,模型获取情感词特征对分类效果影响较大,而识别情感词往往可能不需要很多先验知识,预训练的词向量就可以很好地表达出情感特征。因此,KCNN与本文模型在这两个数据集上的提升并不明显。

图4为短文本概念化的3个实例,其中前两条为新闻分类,第1条短文本分类为商业新闻,第2条分类为科技新闻,而第3条为医学分类,这条短文本被分为寄生虫相关文献。从前两条新闻短文本可以看出,短文本中出现的单词可能使用词向量会表达出错误的语义,例如Apple在短文本中为苹果公司,而词向量很可能将其表达为水果的词义,因此需要引入知识库以获取单词的先验知识,而从知识库中获取的概念集往往也会存在与短文本无关的概念,例如Hip Pop的音乐风格概念与商业新闻无关,所以引入注意力机制,将每个概念与短文本、概念集进行attention计算,把那些与短文本和概念集无关的概念权重减小,防止影响模型分类效果。对于第3条医

学分类的短文本,可以看出短文本中可能会出现某个领域的专业特殊词汇,例如文本中的 *Ascaris lumbricoides* 这一词汇,它们往往在预训练词向量中找不到,因此模型分类能力会变差,而本文模型由于结合知识库,可以获取短文本的先验知识,从而解决这一问题。

Short text: Hip Hop's Online Shop celebrity fashion is booming.

嘻哈的网店名人时尚正在蓬勃发展。

Concepts: music style e commerce venture business  
音乐风格 电子商务企业 商业

Short text: Apple has released the developer version of ios4.2.

苹果发布了ios4.2的开发版本。

Concepts: fruit company software development technical role  
水果 公司 软件开发技术角色

Short text: *Ascaris lumbricoides* is the most common helminth to infect humans.

蛔虫是最常见的人类感染的寄生虫。

Concepts: gastrointestinal nematode long lived parasite  
胃肠道线虫 长期存活的寄生虫

图4 短文本概念化实例

Fig.4 Examples of short text conceptualization

从上文可以看出,本文模型对比 KCNN 模型在各个领域数据集中均有一定的提升。由于本文模型借鉴 Transformer 结构,使用双向 GRU 与多头自注意力相结合,对短文本进行编码,在输入序列内部做 Attention 计算,给重点词增加较高的 Attention 权重,从而获取序列内部之间的联系,同时 multi-Head 部分把每一个由 self-attention 计算出来的 head 使用不同的线性变换,学习出不同的词语关系。本文模型同时使用注意力机制对输入的概念集进行编码,通过计算短文本与其概念集中的概念的注意力权重,赋予与短文本密切相关的概念更高的权重,可以一定程度上解决 KCNN 模型中输入的噪声概念,使重要的概念获取更高的权重,因此文本模型有一定的优越性。

从上文实验结果可以看出,AT-CNN 模型的分类效果全面优于 CNN 模型,由于 CNN 将所有词都同等对待,提取每一处的局部特征,无法判别输入文本的特征词与类别的相关性,没有识别关键字的能力,因此在文本分类任务中表现一般。而融合注意力机制的模型 AT-CNN,提升了网络的特征选择能力,使其在文本分类任务中有更为出色的效果。在 Twitter 的分类任务中,相比于 CNN 模型,准确率提升了 3.4%,在 Movie Review 和 TagMyNews 数据集中,分别提升了 1.7% 和 2.3%,验证了注意力机制的有效性。结合一般形式注意力机制的 AT-LSTM 模型有着较好的表现,AT-LSTM 相比 CNNs-LSTMs 也有较好的提升,在 Ohsumed 数据集中,准确率提升 1.2%,这说明注意力机制在模型训练时可以关注特定目标的特征信息,从而使网络更好地识别文本的类别。

为验证短文本概念集两种注意力机制的有效性,本文研究了调节两种注意力权重的参数  $\gamma$  对本文模型结果的影响,经过手动调节参数  $\gamma$ ,将  $\gamma$  从 0 变为 1,间隔为 0.25,实验结果如表 3 所示。从表 3 可以看出,一般当  $\gamma=0.25$  或  $\gamma=0.50$  时模型效果最佳,具体参数选定需视数据集而定。当参数  $\gamma$  设为 0 或 1 时,模型效果都较差。这是由于当  $\gamma=1$  时,模型忽略了每个概念相对于概念集的重要性,从而导致了模型的性能下降。而当  $\gamma=0$  时,模型忽略了概念相对于短文本之间的语义相似度,在这种情况下会导致与短文本无关的概念可能会被赋予较大的权重,影响模型的分类效果。

表3 参数  $\gamma$  对模型的影响

Table 3 Impact of parameter  $\gamma$  on the model

参数	Twitter	AGNews	Ohsumed	Movie Review	TagMyNews
$\gamma=0.00$	0.669 4	0.719 8	0.393 8	0.776 7	0.624 8
$\gamma=0.25$	0.670 2	0.724 6	0.401 6	0.784 7	0.628 7
$\gamma=0.50$	0.668 9	0.739 5	0.406 9	0.780 1	0.631 0
$\gamma=0.75$	0.665 3	0.730 8	0.396 7	0.778 5	0.627 3
$\gamma=1.00$	0.663 1	0.724 7	0.389 9	0.773 8	0.625 8

### 3 结束语

针对短文本分类任务,本文提出一种融合知识图谱与自注意力机制的 GRU 模型。该模型通过双向 GRU 编码短文本上下文信息,结合自注意力机制使模型关注短文本内部词语的关系,挖掘文本深层次的特征信息。同时使用注意力机制对短文本概念集编码,使模型可以获取短文本的先验知识,解决短文本缺乏上下文信息的问题。在 5 个不同领域的评测数据集上的实验结果验证了本文方法的可行性和有效性。由于本文使用的第三方知识库为微软在 2016 年发布的知识图谱库,可能会出现新词无法查询的问题,因此下一步计划使用 YAGO3 新知识库对模型进行改进。

### 参考文献

- [1] DING Zhaoyun, JIA Yan, ZHOU Bin. Survey if data for microblogs [J]. Journal of Computer Research and Development, 2014, 51(4): 691-706. (in Chinese)  
丁兆云,贾焰,周斌. 微博数据挖掘研究综述[J]. 计算机研究与发展, 2014, 51(4): 691-706.
- [2] WAWRE S V, DESHMUKH S N. Sentiment classification using machine learning techniques[J]. International Journal of Science and Research, 2016, 5(4): 819-821.
- [3] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. [2019-10-20]. <https://arxiv.org/pdf/1301.3781.pdf>.
- [4] KIM Y. Convolutional neural networks for sentence classification[EB/OL]. [2019-10-20]. <https://arxiv.org/pdf/1408.5882.pdf>.



- [ 5 ] KALCHBRENNER N, GREIFENSTETTE E, BLUNSOM P. A convolutional neural network for modelling sentences [C]//Proceedings of the 52nd IEEE Annual Meeting of the Association for Computational Linguistics. Washington D. C. , USA; IEEE Press, 2014: 655-665.
- [ 6 ] ZHANG X, ZHAO J, LECUN Y. Character-level convolutional networks for text classification [C]//Proceedings of IEEE NIPS' 15. Washington D. C. , USA; IEEE Press, 2015: 649-657.
- [ 7 ] TANG Duyu, QIN Bing, LIU Ting. Document modeling with gated recurrent neural network for sentiment classification [C]//Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing. Washington D. C. , USA; IEEE Press, 2015: 1422-1432.
- [ 8 ] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proceedings of IEEE NIPS' 15. Washington D. C. , USA; IEEE Press, 2017: 5998-6008.
- [ 9 ] HUANG B, OU Y, CARLEY K M. Aspect level sentiment classification with attention-over-attention neural networks [C]//Proceedings of International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. Berlin, Germany; Springer, 2018: 197-206.
- [ 10 ] SHI Libin, LIU Jingshuang, ZHANG Xiong. Attention aware bidirectional gated recurrent unit based framework for sentiment analysis [C]//Proceedings of International Conference on Knowledge Science, Engineering and Management. Berlin, Germany; Springer, 2018: 67-78.
- [ 11 ] WANG Zhongyuan, WANG Haixun, WEN Jirong, et al. An inference approach to basic level of categorization [C]//Proceedings of the 24th ACM International Conference on Information and Knowledge Management. New York, USA; ACM Press, 2015: 653-662.
- [ 12 ] WANG Jin, WANG Zhongyuan, ZHANG Dawei, et al. Combining knowledge with deep convolutional neural networks for short text classification [C]//Proceedings of IJCAI' 17. Washington D. C. , USA; IEEE Press, 2017: 2915-2921.
- [ 13 ] LEE J, CHO K, HOFMANN T. Fully character-level neural machine translation without explicit segmentation [J]. Transactions of the Association for Computational Linguistics, 2017, 5(1): 365-378.
- [ 14 ] SUCHANEK F M, KASNECI G, WEUKUM G. Yago: a large ontology from wikipedia and wordnet [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2008, 6(3): 203-217.
- [ 15 ] HU Linmei, YANG Tianchi, SHI Chuan, et al. Heterogeneous graph attention networks for semi-supervised short text classification [C]//Proceedings of Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Washington D. C. , USA; IEEE Press, 2019: 4823-4832.
- [ 16 ] YAO Liang, MAO Chengsheng, LUO Yuan. Graph convolutional networks for text classification [C]//Proceedings of AAAI Conference on Artificial Intelligence. [S. l. ]: AAAI Press, 2019: 7370-7377.
- [ 17 ] VITALE D, FERRAGINA P, SCAIELLA U. Classification of short texts by deploying topical annotations [C]//Proceedings of European Conference on Information Retrieval. Berlin, Germany; Springer, 2012: 376-387.
- [ 18 ] WANG Shengyu, ZENG Biqing, SHANG Qi, et al. Word attention-based convolutional neural networks for sentiment analysis [J]. Journal of Chinese Information Processing, 2018, 32(9): 123-131. (in Chinese)  
王盛玉, 曾碧卿, 商齐. 基于词注意力卷积神经网络模型的情感分析研究 [J]. 中文信息学报, 2018, 32(9): 123-131.
- [ 19 ] CLICHE M. Bb\_twtr at SemEval-2017 task 4: Twitter sentiment analysis with CNNs and ISTMS [EB/OL]. [2019-10-20]. <https://arxiv.org/pdf/1704.06125.pdf>.
- [ 20 ] BAZIOTIS C, PELEKIS N, DOULKERIDIS C. Datastories at semeval-2017 task 4: deep lstm with attention for message-level and topic-based sentiment analysis [C]//Proceedings of the 11th IEEE International Workshop on Semantic Evaluation. Washington D. C. , USA; IEEE Press, 2017: 747-754.