

文章编号: 1003-0077(2021)03-0115-10

## 基于 ELMo 和 Transformer 混合模型的情感分析

赵亚欧<sup>1,2</sup>, 张家重<sup>1</sup>, 李贻斌<sup>3</sup>, 王玉奎<sup>1</sup>

(1. 浪潮集团 金融信息技术有限公司, 山东 济南 250101;

2. 济南大学 信息科学与工程学院, 山东 济南 250022;

3. 山东大学 控制科学与工程学院, 山东 济南 250061)

**摘要:** 针对循环神经网络模型无法直接提取句子的双向语义特征, 以及传统的词嵌入方法无法有效表示一词多义的问题, 该文提出了基于 ELMo 和 Transformer 的混合模型用于情感分类。首先, 该模型利用 ELMo 模型生成词向量。基于双向 LSTM 模型, ELMo 能够在词向量中进一步融入词语所在句子的上下文特征, 并能针对多义词的不同语义生成不同的语义向量。然后, 将得到的 ELMo 词向量输入 Transformer 模型进行情感分类。为了实现分类, 该文修改了 Transformer 的 Encoder 和 Decoder 结构。ELMo 和 Transformer 的混合模型是循环神经网络和自注意力的组合, 两种结构可从不同侧面提取句子的语义特征, 得到的语义信息更加全面、丰富。实验结果表明, 该方法与当前主流方法相比, 在 NLPCC2014 Task2 数据集上分类正确率提高了 3.52%; 在酒店评论的 4 个子数据集上分类正确率分别提高了 0.7%、2%、1.98% 和 1.36%。

**关键词:** 情感分析; ELMo 模型; Transformer 模型; 多头自注意力机制; 自然语言处理

**中图分类号:** TP391

**文献标识码:** A

## Sentiment Analysis Based on Hybrid Model of ELMo and Transformer

ZHAO Yaou<sup>1,2</sup>, ZHANG Jiachong<sup>1</sup>, LI Yibin<sup>3</sup>, WANG Yukui<sup>1</sup>

(1. Inspur Financial Information Technology Company Limited, Jinan, Shandong 250101, China;

2. School of Information Science and Engineering, University of Jinan, Jinan, Shandong 250022, China;

3. School of Control Science and Engineering, Shandong University, Jinan, Shandong 250061, China)

**Abstract:** A hybrid model based on ELMo (Embeddings from Language Models) and Transformer is proposed for sentimental analysis. Firstly, the ELMo model based on bilateral LSTM model is applied to generate word vectors that combine the contexts features and word features, with different vectors for different meanings of a polysemous word. Then, the ELMo vector is input into a Transformer with the encoder and decoder modified for sentiment classification. The hybrid model of ELMo and Transformer with two different network structures can extract the semantic features of sentences from different aspects. The experimental results show that, compared with state-of-the-arts methods, the proposed model improves the accuracy by 3.52% on NLPCC2014 Task2 datasets, by 0.7%, 2%, 1.98% and 1.36% on 4 sub-datasets of hotel reviews respectively.

**Keywords:** sentiment analysis; embeddings from language models; transformer model; multi-heads self-attention mechanism; natural language processing

## 0 概述

情感分析是自然语言处理的一个重要应用领域, 其目的是利用计算机技术分析文本, 获得文字背

后的情感信息。随着互联网技术的发展, 人们的活动越来越多地集中在网络上, 人们通过电商网络获取商品信息, 购买商品并发表评论; 利用新浪、网易等主流网站获取新闻信息, 并发表自己的观点; 利用微博、微信等 APP 与其他人互动。如果能对网络上

收稿日期: 2019-12-03 定稿日期: 2020-03-10

基金项目: 国家重点研发计划云计算和大数据重点专项(2016YFB1001100, 2016YFB1001104); 国家自然科学基金青年项目(61702218)

的文本进行情感分析,则有助于电商了解用户需求,实现对商品的改进;有助于新闻媒体了解用户的喜好,进行新闻的个性化推送;也有助于政府机构对舆情进行监控,避免重大舆情事件的发生。

传统的情感分析通常是采用人工抽取文本特征的方式,然后通过训练机器学习分类器实现文本的情感分类。常用的机器学习分类器有支持向量机(support vector machine, SVM)<sup>[1]</sup>、朴素贝叶斯(naïve bayes, NB)<sup>[2]</sup>、深度森林(deep forest, DF)<sup>[3]</sup>等。但此类方法需要针对不同的应用领域构造不同的特征提取方法,需要相关领域专家参与。

近年来,人工神经网络方法也被用于情感分析。人工神经网络可以自动抽取文本特征,具有人力成本低、领域知识要求少、应用范围广的特点,渐渐成为该领域的主流。Kim<sup>[4]</sup>首先将神经网络方法用于情感分类,该方法将 Word2Vec 工具在 Google News 文本语料上训练所获得的词向量作为神经网络输入;然后利用卷积神经网络(convolutional neural network, CNN)模型对句子的情感倾向进行分类。其后,Attardi 等<sup>[5]</sup>使用卷积神经网络进行情感分类,并在三分类情感数据集上取得了较好的结果。

由于卷积神经网络无法有效提取词语序列之间的顺序信息,Socher 等<sup>[6]</sup>提出使用递归神经网络(recurrent neural network, RNN)进行情感分析,并取得了不错的效果。此外,长短时记忆网络(long-short term memory, LSTM)也被用来进行情感分析<sup>[7]</sup>,相对于递归神经网络,LSTM 能更好地提取序列之间的长程信息,效果更好。

为了解决 LSTM 门限单元计算复杂度高的问题,Cho 等<sup>[8]</sup>提出一种 LSTM 的替代方案——门限循环单元(gated recurrent unit, GRU)。与 LSTM 模型相比,GRU 模型结构更加简单,可以大大提高模型的训练和推理速度。

注意力机制是最近 NLP 领域引入的一个重要概念,其核心是对观察到的数据分配权重,通过权重分配达到提取文本核心语义信息的目的。2016 年, Bahdanau 等<sup>[9]</sup>首先将注意力机制用在机器翻译领域。其后, Luong 等<sup>[10]</sup>、Yin 等<sup>[11]</sup>、Wang 等<sup>[12]</sup>将注意力机制和神经网络结合,进行情感分析。2017 年, Vaswani 等<sup>[13]</sup>提出了多头自注意力(multi-heads self-attention),该机制能够通过对句子本身分配权重、计算加权和,从而从中抽取有效的语义信息。多头自注意力基于 Transformer 模型,最初用于机器翻译, Radford 在此基础上提出了 GPT 模

型<sup>[14]</sup>,该模型仅保留了 Transformer 的 decoder 部分,剔除了 decoder 部分的第二个多头注意力结构。这样结构更加简单,与传统的 RNN 模型相比,运行速度更快、精度更高。

GPT 模型仅是一个前向语言模型,无法利用 token 对应的下文信息。为了解决这个问题,Devlin 等<sup>[15]</sup>提出了 BERT 模型,该模型采用类似完形填空的方式,随机选择句子中的 token 进行 mask,并同时利用 mask 的上下文信息预测该 token。该模型虽然是一个双向模型,但在训练中引入了无关的 mask 标记,影响后续推导。其后, Yang 等<sup>[16]</sup>提出了 XLNet 模型,该模型采用随机排列 token 的方式,并引入了双流注意力机制,不但能有效地利用句子的上下文信息,也避免了 mask 标记的引入。

使用神经网络进行情感分析,除需要选择一个合适的神经网络分类模型之外,获得合适的词向量表示也十分重要。词语的表征最早使用 one-hot、bag-of-words 等离散表示,后来基于神经网络的词嵌入(word embedding)方法被引入用来生成词语的紧致连续表示,这其中的代表有 Word2Vec<sup>[17]</sup>和 GloVe<sup>[18]</sup>等方法。由于此类方法不需要任何先验知识,只要提供文本语料就能训练出有效的语义表示,因此渐渐成为主流。但此类方法缺点也很明显,主要是仅能获得词语的单个语义表示,对于多义词,所获得的词向量是多个语义的合成,这在很大程度上影响了词语语义表示的准确性,给后续任务的使用带来不便。

为了解决这个问题,本文提出了基于 ELMo(embeddings from language models)和 Transformer 的混合模型,用于短文本情感分析。ELMo 是一个基于双向 LSTM 网络的语言模型,通过学习预训练语料,能够得到词语的深度上下文嵌入向量。该向量不仅包含词语本身的语义信息,还包含其对应上下文的语境信息,与传统词嵌入方法相比,优势十分明显。

在情感分类模型方面,本文使用改进的 Transformer 模型,Transformer 模型是经典的 Seq2Seq 模型,使用多头自注意力机制(multi-heads self-attention),并在机器翻译任务中取得了很好的效果。本文对该模型进行了改进,更改了 Transformer 的解码器,使其更加适合情感分类问题。实验在 NLPCC2014 Task2 和谭松波等人的酒店评论数据集上进行,结果表明,本文提出融合 ELMo 和改进 Transformer 的情感分析模型,其效果均高于主

流方法。

## 1 ELMo 模型

ELMo 模型于 2018 年由华盛顿大学的 Peters 等<sup>[19]</sup>提出,目的是建立一个上下文相关的词语向量,为多义词提供更好的表示,克服传统词嵌入模型只能表达词语单一语义的问题。该模型的思路是利用双向 LSTM 网络,通过在预训练语料上训练语言模型,得到词语上下文相关的语义向量。

### 1.1 利用 LSTM 网络构建语言模型

语言模型是计算给定句子出现概率的模型。假设含有  $N$  个词语的句子  $S = \{t_1, t_2, \dots, t_N\}$ ,如果句子中第  $k$  个词语  $t_k$  出现的概率仅与其前面出现的  $k-1$  个词有关,则该语言模型为前向语言模型。同理,如果假设句子中第  $k$  个词语  $t_k$  出现的概率仅与其后面  $N-k$  个词语有关,则该语言模型为后向语言模型。

ELMo 模型利用 LSTM 网络构建前向语言模型。假设  $t_1, t_2, \dots, t_k$  对应的词向量分别为  $x_1, x_2, \dots, x_k$  (该词向量可使用 Word2Vec 等工具计算

获得),将其依次输入一个含有  $L$  层的 LSTM 网络,得到对应层的隐状态  $h_{l,1}, h_{l,2}, \dots, h_{l,k}$ ,其中  $l$  为 LSTM 的层数  $l = \{1, 2, \dots, L\}$ 。取最后一层隐状态  $h_{L,1}, h_{L,2}, \dots, h_{L,k}$ ,输入 softmax 层转化为概率输出  $y_{L,1}, y_{L,2}, \dots, y_{L,k}$ ,则该输出为前向语言模型对应词语的概率分布  $P(t_1), P(t_2 | t_1), \dots, P(t_k | t_1, t_2, \dots, t_{k-1})$ 。

同理,ELMo 利用另外一个 LSTM 网络构建后向语言模型,该 LSTM 的方向和前向 LSTM 的方向相反。将  $t_1, t_2, \dots, t_k$  对应的词向量  $x_1, x_2, \dots, x_k$  输入,同样得到对应层的隐状态  $h'_{l,1}, h'_{l,2}, \dots, h'_{l,k}$ ,将最终层对应的隐状态经过 softmax 映射转换为概率输出  $y'_{L,1}, y'_{L,2}, \dots, y'_{L,N}$ ,此概率为后向语言模型对应词语的概率分布  $P(t_1 | t_2, \dots, t_N), P(t_2 | t_3, \dots, t_N), \dots, P(t_k | t_{k+1}, \dots, t_N)$ 。

最后,为了构建双向语言模型,连接两个方向 LSTM 最终层的隐状态,得到  $H_{L,1}, H_{L,2}, \dots, H_{L,N}$ ,其中  $H_{L,k} = \{h_{L,k}, h'_{L,k}\}$ , $k$  表示第  $k$  个位置词语输入对应的隐状态。 $H_{L,k}$  经过 softmax 激励函数得到  $Y_k$ ,即为双向语言模型第  $k$  个词语对应的概率分布  $P(t_k | t_1, t_2, \dots, t_{k-1}, t_{k+1}, \dots, t_N)$ 。模型如图 1 所示。

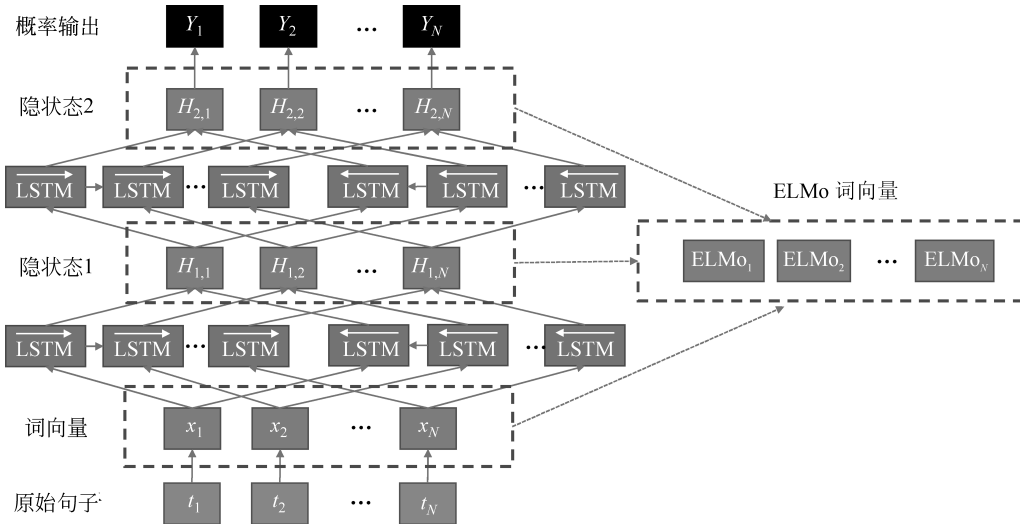


图 1 ELMo 模型

### 1.2 ELMo 词向量

ELMo 词向量通过对双向 LSTM 每一层的隐状态向量加权求和获得。假设前向 LSTM 和后向 LSTM 隐层连接后的向量为  $H_{l,k}$ ,输入词向量为  $x_k$ ,则 ELMo 词向量表示为:

$$\text{ELMo}_k = \gamma \left( s_0 x_k + \sum_{j=1}^L s_j H_{j,k} \right) \quad (1)$$

其中  $\gamma$  为缩放因子,  $s_j$  为针对隐状态的归一化系数,表示每一层隐状态的占比。这两组参数通过后续任务优化获得。如果不对后续任务进行二次优化,也可以使用 LSTM 最后一层隐状态作为 ELMo 词向量,即

$$\text{ELM}_{O_k} = \mathbf{H}_{L,k} \quad (2)$$

## 2 基于 Transformer 的情感分类模型

### 2.1 自注意力机制

注意力机制(attention mechanism)来源于人类视觉处理过程,最初应用在机器翻译任务中。其核心思想是,通过计算译文句中单词与原文句中单词之间的相互关系,得到译文单词相对于原文单词的权重分布,然后通过加权处理,得到译文单词的最佳语义表示。自注意力是注意力机制的一种特殊形式,也叫内部注意力,特指注意力的计算仅对于同一句子进行。

自注意力机制通过三个矩阵实现,分别是查询(Query)矩阵  $\mathbf{Q}$ 、键(Key)矩阵  $\mathbf{K}$  和值(Value)矩阵  $\mathbf{V}$ 。 $\mathbf{Q}$  和  $\mathbf{K}$  用于计算输入词语与句子中其他词语之间的相似性,并根据相似性计算注意力的分配比例,值矩阵  $\mathbf{V}$  表示句中词语对应的注意力值。 $\mathbf{Q}$ 、 $\mathbf{K}$ 、 $\mathbf{V}$  可以采用多组,每一组被称为一个头(head),此时被称为多头自注意力。

实际计算中,三个矩阵均为句子矩阵  $\mathbf{X}$ ,  $\mathbf{X} \in N \times d$  的线性变换,其中  $N$  为句子长度, $d$  表示词向量的维度,其计算如式(3)所示。

$$\begin{aligned} \mathbf{Q}_i &= \mathbf{XW}_Q^i \\ \mathbf{K}_i &= \mathbf{XW}_K^i \\ \mathbf{V}_i &= \mathbf{XW}_V^i \end{aligned} \quad (3)$$

每一组  $\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i$  矩阵计算一个头的注意力数值  $\text{head}_i$ ,将多个  $\text{head}_i$  连接,得到最终的多头注意力向量  $\mathbf{Z}$ ,计算如式(4)所示。

$$\begin{aligned} \text{head}_i &= \text{attention}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) \\ \mathbf{Z} &= \text{concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_m) \end{aligned} \quad (4)$$

### 2.2 Transformer 模型

Transformer 由 Google 的 Vaswani 等人于 2017 年提出,最早用于自然语言翻译。该模型为典型的 Encoder-Decoder 结构,Encoder 和 Decoder 均可叠加多层。Encoder 模块包含一个自注意力层和一个前馈神经网络层。自注意力层采用多头自注意力机制,前馈神经网络是一个标准的两层神经网络,第一层使用 Relu 激活函数,第二层无激活。此外,注意力层和前馈神经网络都采用残差连接,并进行批标准化(batch normalization),进一步增强网络的泛化能力。

Decoder 模块包含两个自注意力层和一个前馈神经网络,实现译文句子的解码。该模块的第一个注意力层为带屏蔽的多头自注意力层(masked multi-head attention),目的是在推导过程中屏蔽未来输入。该模块的第二个注意力层和前馈神经网络与 Encoder 模块中的结构几乎完全一样,唯一的区别是,Decoder 的输入除了包含前一层的输出之外,还增加了 Encoder 的输出。多个 Decoder 模块可以叠加,最后一个 Decoder 的输出经过一个线性变换,并利用 softmax 函数,得到输出词语的预测概率。

Transformer 本质上是一个自编码器,不能利用词语之间的顺序信息,所以引入位置嵌入向量(position embedding, PE)来表示词语的位置。PE 的计算如式(5)所示。

$$\begin{aligned} \mathbf{PE}_{(\text{pos}, 2i)} &= \sin(\text{pos}/10000^{2i/d}) \\ \mathbf{PE}_{(\text{pos}, 2i+1)} &= \cos(\text{pos}/10000^{(2i+1)/d}) \end{aligned} \quad (5)$$

其中  $d$  表示 PE 的维数,  $\text{pos}$  表示词语在句中的位置。 $2i$  表示 PE 的偶数维度,  $2i+1$  表示奇数维度。Transformer 实际使用的输入是词语词向量和 PE 之和。

Transformer 模型抛弃了传统的 RNN 网络结构,全部使用自注意力机制实现。该模型不像 RNN 模型那样每次只能提取一个方向的特征,可直接提取句子的双向语义特征。此外,该模型还避免了 RNN 的梯度衰减和长程信息丢失问题,且更易于并行实现。

### 2.3 改进的 Transformer 的分类模型

虽然 Transformer 模型有诸多好处,但该模型是经典的 Seq2Seq 模型,输入、输出都是词语序列,无法直接用于情感分类任务。为了解决这个问题,本文对 Transformer 模型进行了修改,主要有以下几点:

(1) Transformer 的 Encoder 模块的作用是将整个句子的语义融合到每个词语中,这样做既可以丰富词语的语义,也有利于多义词的语义消歧。但由于多头注意力运算是针对单个词向量进行的,因此编码结果仍然是单个词向量。然而,情感分析需要的是句子的整体语义表示,需要将编码后的词向量进行融合。为解决这个问题,本文在最后一个 Encoder 模块中增加了一个 concat 单元,将所有的词语向量进行连接,构成整个句子向量。假设句子的编码器的输出为  $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_N$ , 每个向量的维度



为  $d$ , 则连接之后的句子向量为  $\mathbf{E} = \{e_1, e_2, \dots, e_N\}$ ,  $\mathbf{E} \in R^{N \times d}$ 。

(2) 对于 Transformer 的 Decoder 模块, 其主要目的是根据每个词语的 Encoder 信息, 结合译文上文信息, 对译文进行解码, 这部分对分类任务而言是不必要的。本文去除了原结构的两个自注意力层, 仅保留残差连接和前馈网络。其考虑是, concat 模块的作用只是单纯地连接多个 Encoder 输出的词向量, 并不能很好地对其语义特征进行融合, 因此, 保留一个前馈网络, 将连接后的向量进行二次映射, 以保证更好的融合效果。相对于原结构, 前馈神经网络不再针对单个词语进行, 而是针对整个句子向量  $\mathbf{E}$  进行; 保留残差连接目的是考虑到 Decoder 模块可以叠加多层, 残差连接能保证初始信息向更深层模块传递。

Decoder 模块的输入有两个, 一个是 Encoder 编码后的句子向量  $\mathbf{E}$ , 另一个是前一层 Decoder 的输出  $\mathbf{E}'$ 。对于第一层的 Decoder 结构, 其前一层 Decoder 输出  $\mathbf{E}'$  为原始句子词向量连接后的向量  $\mathbf{X} = \{x_1, x_2, \dots, x_N\}$ ,  $\mathbf{X} \in R^{N \times d}$ 。

(3) 最终层 Decoder 的输出为  $\mathbf{Z}$ ,  $\mathbf{Z} \in R^{N \times d}$ 。为了进一步抽取主要特征, 添加一个 max 池化层, 获取融合后词向量在所有维度上的最大值, 其输出为  $\mathbf{Z}'$ ,  $\mathbf{Z}' \in R^d$ , 计算公式如式(6)所示。

$$\mathbf{Z}'_d = \max_i \mathbf{Z}_{i,d} \quad (6)$$

其中,  $\mathbf{Z}'_d$  表示输出词向量第  $d$  维分量值,  $\mathbf{Z}_{i,d}$  表示  $\mathbf{Z}$  矩阵  $i$  行  $d$  列的分量值,  $\max$  表示取  $\mathbf{Z}$  矩阵每一行的最大值。

最后, 池化后输出  $\mathbf{Z}'$  接一个线性映射单元, 并利用 softmax 函数计算各类情感倾向的概率, 计算公式如式(7)所示。

$$\mathbf{Y} = \text{softmax}(\mathbf{Z}'\mathbf{W} + \mathbf{b}) \quad (7)$$

其中,  $\mathbf{W}, \mathbf{W} \in R^{d \times 2}$  为神经网络权值矩阵,  $\mathbf{b}, \mathbf{b} \in R^2$  为网络偏置项。  $\mathbf{Y}$  表示二分类输出, 表示输入文本属于正面、负面情感的概率。改进后的模型结构如图 2 所示。

模型优化采用交叉熵损失函数, 如式(8)所示。

$$l = - \sum_{i=1}^S \sum_{k=1}^2 \mathbf{Y} p_{i,k} \log(\mathbf{Y}_{i,k}) \quad (8)$$

其中,  $S$  表示训练样本总数,  $k$  表示类别,  $\mathbf{Y} p_{i,k}$  表示第  $i$  个样本对应第  $k$  个类别的期望输出概率,  $\mathbf{Y}_{i,k}$  表示第  $i$  个样本对应第  $k$  个类别的模型实际输

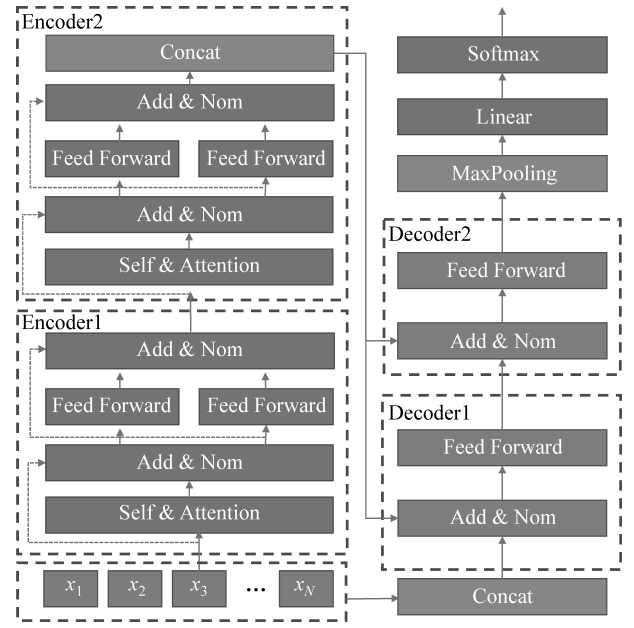


图 2 改进的 Transformer 分类模型

出概率。

## 2.4 基于 ELMo 和 Transformer 的混合模型

基于上述模型, 本文提出了基于 ELMo 和 Transformer 的混合模型用于情感分析, 该模型可分为 6 个主要模块, 结构图如图 3 所示。

(1) 句子分词模块。利用中文切词工具, 如结巴分词、StanfordNLP 分词工具对中文句子进行分词, 将其切分为独立的 token。

(2) 词向量计算模块。利用 Word2Vec 工具, 在预训练语料中训练, 获得词嵌入向量。然后将上一步获得的 token 转换为对应词向量。

(3) ELMo 词向量计算模块。将 Word2Vec 获得的词向量输入 ELMo 模型, 并在预训练语料中训练该模型, 取双向 LSTM 最后一层的隐状态向量并进行连接, 生成 ELMo 词向量。

(4) Encoder 计算模块。ELMo 词向量输入 Transformer 的 Encoder 模块, 获得对应的编码向量。Encoder 模块可以叠加多个, 进行多次编码。

(5) Decoder 计算模块。连接编码向量并输入 Decoder 模型进行解码, 经过多次特征映射, 获得整个句子的语义向量表示。Decoder 模块同样可以叠加多个, 进行多次映射。

(6) 分类输出模块。将 Decoder 生成的语义向量输入线性模块, 然后经过 softmax 激励函数, 得到最终的分类结果。

相对于传统的情感分析模型,该模型有如下优点:

(1) 输入分类器的词向量不再是原始的词嵌入向量,而是 ELMo 向量。ELMo 向量既包含词语本身语义,也包含词语对应的上下文语义,包含的信息更加丰富。

(2) 情感分析模型大多基于自回归结构或自编码结构,本文模型采用这两种结构的组合,即 ELMo 的自回归结构(BiLSTM)和 Transformer 的编码器结构(multi-heads attention),两种结构的组合能更好地提取特征。

(3) 本文提出的改进 Transformer 结构利用句中所有词语融合后的向量作为最后的句子向量(Concat+Feed Forward+Max Pooling),与 BERT 仅采用[CLS]标记相比,能更好地表示整个句子。

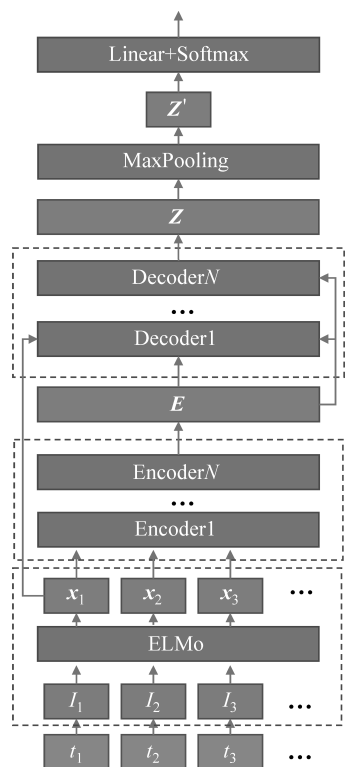


图3 基于 ELMo 和 Transformer 的混合模型结构图

### 3 实验

#### 3.1 数据集

实验在两个数据集上进行,一个是自然语言处理和中文计算国际会议公布的深度学习情感分类数据集(NLPCC2014 Task2)<sup>[20]</sup>。该数据集分为训练集和测试集,其中,训练集包含 10 000 条产品评论

信息,正面和负面评论各 5 000 条;测试集包含 2 500 条产品评论信息,正面和负面评论各 1 250 条。为了评估模型参数,取训练集的 80% 作为实际的训练集,剩余 20% 作为开发集。

另一个是谭松波等<sup>[21]</sup>收集的酒店评论数据集,该数据集包含 10 000 条评论信息,其中正面评论 7 000 条,负面评论 3 000 条。根据不同规模,该数据集被划分为四个不同的数据集,分别是 htl-2 000, htl-4 000, htl-6 000, htl-10 000。其中,前三个数据集为平衡数据集,正面评论和负面评论数据各半,最后一个数据集为非平衡数据集。实验中,为了方便和前人的结果进行比较,采用 10 折交叉验证,每次取 90% 样本作为训练集,剩余 10% 的样本作为测试集。模型参数与 NLPCC2014 数据集的最优参数保持一致,因此不再划开发集。

#### 3.2 实验参数设置

本实验模型采用 ELMo 和 Transformer 的混合模型,其中,ELMo 模型输入词向量维度为 256,双向 LSTM 的层数为 2,展开深度为 30,输出 ELMo 词向量的维度为 512。ELMo 模型采用预训练方式,预训练在百科类问答语料上进行,该语料包含 150 万个预先过滤的问题和答案。<sup>①</sup>

Transformer 的 Encoder 采用多头注意力机制,多头数目定为 8,多头注意力输出向量的维度为 512,神经网络的激励函数采用 gelu(Gaussian Error Linear Units)<sup>[22]</sup>,其形式为:

$$\text{gelu}(x) = 0.5x \left( 1 + \tanh \left[ \sqrt{\frac{2}{\pi}} (x + 0.044715x^2) \right] \right) \quad (9)$$

Decoder 模块采用本文 2.3 节描述的改进模型。

实验评测采用正确率指标(Accuracy, A),其计算如式(10)所示。

$$A = \frac{TP + TN}{TP + FP + TN + FN} \quad (10)$$

其中,TP、TN、FP、FN 分别对应阳性、阴性、假阳性和假阴性样本的数目。

实验中,将整个微博评论短文本视为一个长句子,预处理去除停用词、标点并进行分词,然后输入 ELMo 模型获得该句子对应的 ELMo 词向量集合。输入 Transformer 时,要设定最大句子长度(ELMo

<sup>①</sup> 语料下载地址: [https://github.com/brightmart/nlp\\_chinese\\_corpus](https://github.com/brightmart/nlp_chinese_corpus)

词向量集合所包含最大词向量个数),对长度大于该数值的句子进行截断,对长度小于该数值的句子进行填充。(填充<end>标记对应的 ELMo 向量)。

为了获取最优实验参数,句子最大长度采用 64、128 和 256,Encoder 和 Decoder 层数采用 1~8 层,在 NLPCC2014 Task2 数据集上进行实验,开发集上的结果如图 4 所示。

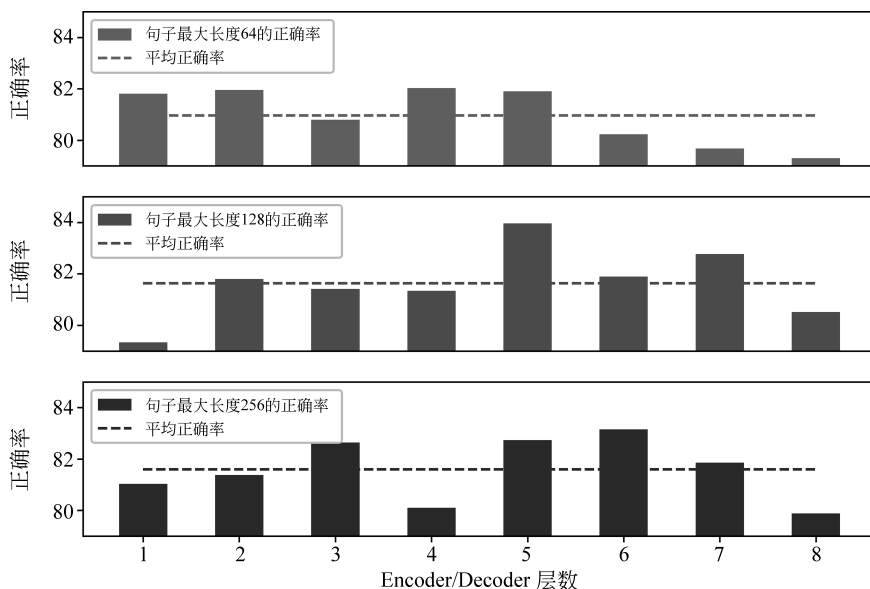


图4 采用不同 Encoder/Decoder 层数和不同句子最大长度在开发集上的实验结果

从图4中可以看出,当句子最大长度为64时,平均正确率最低,为80.96%,句子最大长度为128和256时,正确率相当,但句子最大长度为128时,正确率略高且更稳定。考虑到模型的计算复杂度,选择最佳句子长度为128。对于层数,从图中可以看出,当Transformer层数为6时,效果最好,达到了83.97%,实验结果和原始的Transformer构造的机器翻译模型设置的最优层数相同。

### 3.3 本文模型与其他模型的对比

为了证明本文模型的有效性,实验分别在NLPCC2014 Task2和谭松波的酒店评论数据集上进行,对比实验中采用如下模型:

(1) 经典的LSTM模型。使用Word2Vec提取词语向量,然后利用LSTM模型进行情感分类。

(2) GRU模型。GRU模型是LSTM模型的一种变体,结合了遗忘门和输入门,合成为一个更新门限,相对于LSTM单元更加简单,更利于训练。

(3) CNN模型。输入使用Word2Vec提取词语向量,然后利用卷积、池化操作进一步抽取整个句子的语义特征,最后通过softmax激励进行情感分类。

(4) LSTM+CNN2+CNN3混合模型<sup>[20]</sup>。由

于CNN不考虑句子中词语之间的顺序关系,因此考虑在此基础上融合LSTM模型,进一步丰富所提取的特征,其中CNN2、CNN3分别表示采用卷积核大小为2、3的卷积神经网络。

(5) SVM模型,经典的机器学习模型,其核心思想是找到高维空间中的最优决策面。

(6) NB模型,朴素贝叶斯模型。

(7) BCWCNN模型<sup>[23]</sup>,双线性字词卷积神经网络。为了克服词向量表征语义不充分的问题,选取词向量和字向量作为特征输入卷积神经网络进行训练,然后使用融合后的特征进行分类。

(8) WCTAT-Bi-LSTM模型<sup>[24]</sup>,基于Bi-LSTM的字、词和词性注意力模型。该模型输入采用词向量和字向量的融合向量,分类模型使用含有注意力机制的双向LSTM模型。

(9) BFDF模型<sup>[25]</sup>,基于强化表征学习的深度森林模型。

(10) Google的BERT-base模型,中文模型采用科大讯飞提供的BERT-base预训练模型<sup>[26]</sup>,输入采用字向量。

(11) Word2Vec+Transformer模型。输入采用Word2Vec提取的词向量,分类模型采用本文2.3节介绍的Transformer模型。

(12) ELMo + Transformer。输入向量采用 ELMo 词向量,分类模型采用本文 2.3 节介绍的 Transformer 模型。

上述神经网络模型中,模型(10)、(11)、(12)的 Transformer 采用 gelu 激励函数,其余模型均采用 relu 激励,模型(11)中的 ELMo 也采用 relu 激励。

### 3.4 结果分析

在 NLPCC2014 Task2 数据集上的实验结果如表 1 所示。从表中可以看出,本文提出的基于 ELMo 和 Transformer 的混合模型取得了最好的效果,正确率达到了 79.68%,与 LSTM 模型相比,正确率提升 13.04%;与 CNN 模型相比,正确率提升 6.32%;与 LSTM + CNN2 + CNN3 的融合模型相比,正确率提升 3.52%;与 BERT-base 模型相比,正确率也略有提升。后三个模型均采用 Transformer 作为分类器,其正确率好于其余非 Transformer 模型,这说明多头注意力机制在抽取文本整体语义特征方面,较传统的 CNN 或 RNN 模型更具优势。

表 1 不同方法在 NLPCC2014 Task2 数据集上的实验结果

Method	Accuracy
LSTM <sup>[20]</sup>	0.666 4
GRU <sup>[20]</sup>	0.739 2
CNN <sup>[20]</sup>	0.733 6
LSTM + CNN2 + CNN3 <sup>[20]</sup>	0.761 6
BERT <sup>[26]</sup>	0.793 8
Word2Vec + Transformer	0.764 4
ELMo + Transformer	<b>0.796 8</b>

本文方法使用 ELMo 词向量作为 Transformer 的模型输入,与使用 Word2Vec 词向量相比,分类性能提升了 3.24%,这说明 ELMo 中的双向 LSTM 网络在提取词向量特征方面作用明显。此外,ELMo 是预训练模型,训练采用了超大规模语料,相对于仅使用任务数据进行训练双向 LSTM 模型,其抽取的特征涵盖的范围更广,更具一般性。

在酒店评论数据集上,实验分别对 htl-2 000, htl-4 000, htl-6 000 和 htl-10 000 四个数据集进行了测试,从表 2 可以看出,相较于 BCWCNN 模型,正确率分别提升了 1.2%、0.25%、1.48 和 1.96%,对于采用融合字词向量特征的双向注意力 LSTM 模型,正确率分别提升了 0.7%、2%、1.98%和 1.36%。这说明本文方法相对于传统方法具有优越性,这一

方面是由于 ELMo 向量相对于传统的词向量融合了词语上下文,能够更好地表示多义词,另一方面也是因为 Transformer 的多头注意力机制能够有侧重地对多个词向量进行融合,更好地提取整句语义。

表 2 不同方法在酒店评论数据集上的实验结果

Method	Accuracy			
	htl-2 000	htl-4 000	htl-6 000	htl-10 000
SVM <sup>[23]</sup>	0.7325	0.826 0	0.830 0	0.848 0
NB <sup>[23]</sup>	0.677 6	0.684 4	0.696 0	0.684 0
CNN <sup>[23]</sup>	0.871 8	0.883 6	0.912 8	0.928 0
BCWCNN <sup>[23]</sup>	0.930 0	0.940 0	0.920 0	0.917 0
WCTAT-Bi-LSTM <sup>[24]</sup>	0.935 0	0.922 5	0.915 0	0.923 0
BDFD <sup>[25]</sup>	—	—	—	0.918 1
BERT <sup>[26]</sup>	0.940 0	<b>0.947 5</b>	0.921 7	0.932 0
Word2Vec + Transformer	0.922 0	0.919 3	0.908 2	0.910 0
ELMo + Transformer	<b>0.942 0</b>	0.942 5	<b>0.934 8</b>	<b>0.936 6</b>

与 BERT 模型相比,除在 htl-4 000 数据集上本文模型正确率略低之外,在另外三个数据集上正确率分别提升 0.2%、1.31%和 0.46%。这主要是因为 BERT 模型仅使用输出的第一个[CLS]标记作为整个句子的语义向量,而本文提出改进的 Transformer 模型将所有词语的词向量进行 Concat 并在 Decoder 阶段进行二次映射,最后通过 Max Pooling 抽取融合后词向量在每个维度的最大值,将该值作为整个句子的语义向量进行预测,该向量所包含的语义更加全面。

此外,本文方法与 BERT 等方法的预训练阶段不同,BERT 直接预训练分类模型,而本文方法预训练的是 ELMo 模型,没有预训练基于 Transformer 的分类模型。这样,无须引入额外 mask 标记,减少了 mask 对后续推导的影响。

## 4 结束语

本文提出了基于 ELMo 和 Transformer 的混合模型用于情感分析,该模型抛弃了传统的词嵌入方法(如 Word2Vec、GloVe),利用 ELMo 模型抽取词向量。与传统方法相比,ELMo 模型引入双向 LSTM 模型学习词语的上下文,进一步丰富了词向



量的语义,能更好地表示多义词。分类器采用 Transformer 模型,采用多头注意力机制,能够以不同的方式对句子中的词向量进行融合,有侧重地抽取句子的整体语义。Transformer 相对于 RNN 模型,是真正意义上的双向模型,能够直接融合词语的上下文特征,而非简单连接两个方向提取的特征向量,因此提取的特征更加准确。实验在 NLPCC2014 Task2 情感分类数据集和谭松波的酒店评论数据集上进行,结果证明了本文方法的有效性。

近些年来,注意力机制被证明在提取语义特征方面十分有效。对于 ELMo 模型,本文采用双向 LSTM 网络生成上下文相关的词向量,并没有引入注意力机制,下一步考虑在模型中引入注意力机制,进一步增强 ELMo 模型表征词向量的能力。此外,先在大规模语料中预训练模型,然后在新的任务上精调参数已成为当前自然语言处理的主流范式,此时,预训练的语言模型也是后续任务的分类模型,其中代表如 BERT。因此,构造基于 ELMo 的预训练和分类综合模型也是下一步的研究方向。

## 参考文献

- [1] 李婷婷,姬东鸿. 基于 SVM 和 CRF 多特征组合的微博情感分析[J]. 计算机应用研究, 2015, 32(4): 978-981.
- [2] 苏莹,张勇,胡珀,等. 基于朴素贝叶斯与潜在狄利克雷分布相结合的情感分析[J]. 计算机应用, 2016, 36(6): 1613-1618.
- [3] Zhou Z H, Feng J. Deep Forest: Towards an alternative to deep neural networks[J]. arXiv preprint arXiv: 1702.00883v1, 2017.
- [4] Kim Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. New York: ACM, 2014: 1746-1751.
- [5] Attardi G, Sartiano D. UniPI at SemEval-2016 Task 4: Convolutional neural networks for sentiment classification [C]//Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016: 220-224.
- [6] Socher R, Pennington J, Huang E H, et al. Semi-supervised recursive autoencoders for predicting sentiment distributions[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 151-161.
- [7] Wang X, Liu Y, Sun C J, et al. Predicting polarities of tweets by composing word embeddings with long short-term memory[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, 1: 1343-1353.
- [8] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J/OL]. arXiv preprint arXiv: 1406.1078, 2014.
- [9] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J/OL]. arXiv preprint arXiv: 1409.0473, 2014.
- [10] Luong M T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation[J/OL]. arXiv preprint arXiv: 1508.04025, 2015.
- [11] Yin W, Schutze H, Xiang B, et al. ABCNN: Attention-based convolutional neural network for modeling sentence pairs[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 259-272.
- [12] Wang Y, Huang M, Zhu X, et al. Attention-based LSTM for aspect-level sentiment classification[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016: 606-615.
- [13] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems, 2017: 5998-6008.
- [14] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[EB/OL].[2019-12-03][https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language\\_understanding\\_paper.pdf](https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language_understanding_paper.pdf).
- [15] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J/OL]. arXiv preprint arXiv: 1810.04805, 2018.
- [16] Yang Z, Dai Z, Yang Y, et al. XLNET: Generalized autoregressive pretraining for language understanding [C]//Proceedings of the 33rd Conference on Neural Information Processing Systems, 2019: 5754-5764.
- [17] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J/OL]. arXiv preprint arXiv: 1301.3781, 2013.
- [18] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014: 1532-1543.
- [19] Peters M E, Neumann M, Iyyer M, et al. Deep cont-

- extualized word representations[J/OL]. arXiv preprint arXiv: 1802.05365, 2018.
- [20] 杜永萍, 赵晓铮, 裴兵兵. 基于 CNN-LSTM 模型的短文本情感分类[J]. 北京工业大学学报, 2019, 45(7): 48-56.
- [21] Tan S B, Zhang J. An empirical study of sentiment analysis for Chinese documents[J]. Expert Systems with applications, 2008, 34(4): 2622-2629.
- [22] Hendrycks D, Gimpel K. Gaussian error linear units (gelus)[J/OL]. arXiv preprint arXiv: 1606.08415, 2016.
- [23] Wang X, Li J, Yang X, et al. Chinese text sentiment analysis using bilinear character-word convolutional neural networks[C]//Proceedings of the 2017 International of Conference on Computer Science and Application Engineering. Pennsylvania: DEStech Publications, 2017: 36-43.
- [24] 赵富, 杨洋, 蒋瑞, 等. 融合词性的双注意力 Bi-LSTM 情感分析[J]. 计算机应用, 2018, 38(S2): 103-106.
- [25] 韩慧, 王黎明, 柴玉梅, 等. 基于强化表征学习深度森林的文本情感分类[J]. 计算机科学, 2019, 46(7): 172-178.
- [26] Cui Y, Che W, Liu T, et al. Pre-training with whole word masking for Chinese BERT[J/OL]. arXiv preprint arXiv: 1906.08101, 2019.



赵亚欧(1982—), 通信作者, 博士, 讲师, 主要研究领域为自然语言处理、人工智能。

E-mail: zhaoyaou@inspur.com



李贻斌(1960—), 博士, 教授, 主要研究领域为机器人、人机交互。

E-mail: liyb@sdu.edu.cn



张家重(1965—), 博士, 教授, 主要研究领域为人工智能、数据挖掘。

E-mail: zhangjzh@inspur.com