文章编号：2096-1618(2022)04-0379-07

基于 BERT-VGG16 的多模态情感分析模型

陈宏松， 安俊秀， 陶全桧， 周 俊
(成都信息工程大学软件工程学院，四川 成都 610225)

摘要：针对传统情感分析方法只采用文本数据无法充分挖掘情感信息，且单模态数据包含的信息量有限，不能很好反映真实情感状态等问题，提出一种引入注意力机制的多模态情感分析模型。首先，该模型使用预训练模型 BERT 和 VGG16 分别从文本数据和图像数据中提取特征。其次，为提高各模态重要特征权重，特征融合时引入注意力机制，融合后的模型可大幅提升数据信息量。实验结果表明，使用基于 BERT-VGG16 引入注意力机制的多模态特征融合模型比单模态和其他多模态特征融合模型在情感分析效果上有显著提升。

关 键 词：情感分析；多模态；BERT-VGG16 模型；注意力机制
中图分类号：TP391.1 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2022.04.003

0 引言

情感是人对客观事物所持有的一种态度，也是反映人类拥有智能的一种表现。在进行交流的过程中，一段文字、一句语音、一张图片就能够判断目标对象在特定场景下的情感状态及对相关事物的态度。情感分析(sentiment analysis)就是利用计算机通过数据和算法模型，分析目标对象准确的情感状态，是一种常见的自然语言处理(NLP)方法的应用[1]。在人的表达和交互过程中，准确地把握相关方的情感状况和态度可以极大地提升人工智能产品的体验，在质检、交互、风控、舆论监督等方面都有着重要的应用。在日常生活中，文字是必不可少的一种通信交流方式，基于情感词典[2]的情感分析方法能够有效地对文本内容的情感进行划分，这种分类方法的关键在于情感词典的覆盖范围。但随着通信技术的不断进步，微博、B 站、抖音等社交平台的普及也使文本数据和数据类型越来越丰富，新的情感词也越来越多，情感词典的维护也越困难。自深度学习蓬勃发展以来，基于深度学习的情感分析方法被研究者们所关注[3]。

Zhang 等[4]提出的 TextCNN 一种基于卷积神经网络(convolutional neural networks，CNN)的文本情感分析模型，这种模型网络结构简单，通过引入词向量计算文本间情感的关系。然而此模型难以评估每个情感特征的重要度，容易忽略文本中的重要情感特征。Zhou 等[5]用长短期记忆网络(long short term memory，LSTM)作为基础模型，提出 C-LSTM 模型，解决了情感分析任务中存在的长期依赖问题。Karim Ahmed 等[6]提出新型网络结构 Transformer，该模型引入自注意力机制(self-attention mechanism)，对输入序列的每个标记进行并行评估，消除了循环神经网络(recurrent neural networks，RNN)中的顺序依赖。这些模型和方法对于单模态文本数据都有着很好的分类效果。

但是单模态文本数据中包含的信息有限，在某些情况下单凭目标文本难以准确判断目标的情感状态[7]。一个极端例子是反讽，反讽往往结合中性或者积极的文本内容和与内容不匹配的图片完成一个与文本内容相反的情感表达。如，“今天天气真好!”这个句子单从文本数据上分析是表示积极的情感，但当配上一张下雨天气的图片时，整体目标情感可以看成是带有反讽意味的消极情感，这种情况仅靠单模态数据很难从根本上解决问题。此外，单模态模型很容易受目标噪声的影响。

为此，本文同时对文本数据和图像数据进行特征提取，提取两个模态的特征向量，通过加入注意力机制，对高权重特征数据向量进行特征融合，得到文本和图像融合的向量表示，经过归一化，能够得到不同情感状态的概率分布，从而得到多模态融合模型。与单模态模型和现有多模态模型相比，本文所构建的 BERT-VGG16 多模态模型在情感分析任务的准确度上有很大的提升。

1 多模态情感分析

多模态情感分析是 NLP 的一个新兴领域，旨在对两个或两个以上的模态数据(如文本、图像、声学注释和面部表情等)同时进行建模，通过跨模态的方式将不同模态信息相互作用，从而增加整体目标任务的准

收稿日期：2021-12-20
基金项目：国家自然科学基金资助项目(71673032)

确性[8]。在数据信息的挖掘方面,多模态情感分析方法通常优于其他单模态模型。因此,研究者们提出了各种多模态融合模型和方法,Zixuan Peng 等[9]结合 CNN 和 RNN 对音频和文本进行融合,提出 CRNN 模型对音频和文本序列中的信息进行编码,比关注单独音频特性的模型能更全面地利用数据中的信息。Vasco Lopes 等[10]使用自动机器学习(automatic machine learning,AutoML)构建随机搜索融合模型,减少了人工对数据的干预。Minping 等[11]结合 LSTM 模型提出情感词感知融合网络(sentimental words aware fusion network,SWAFN),提高了多模态情感词的感知融合表示。然而,不同模态由于数据类型的疏密程度不同,会产生不同的上下文内容,直接融合的方法会损失各模态数据的重要信息[12]。情感本身是一种复杂的信息表示,在情感分析建模中使用不同的模态信息有不同的处理方法和相应的挑战,于是研究者们对不同模态的模型和融合方式进行了研究。

本文提出一种结合文本与图像的多模态融合模型,在文本模态数据中采用 BERT 作为文本预处理模型,在图像模态中采用 VGG16 来提取图像的数据特征。在特征融合前,分别对两个模态数据加入注意力机制(attention echanism),获取特征权重更高的数据集再进行融合,最后再通过 SoftMax 函数进行归一化,得到最终的情感状态。基于 BERT-VGG16 融合模型的结构如图 1 所示。

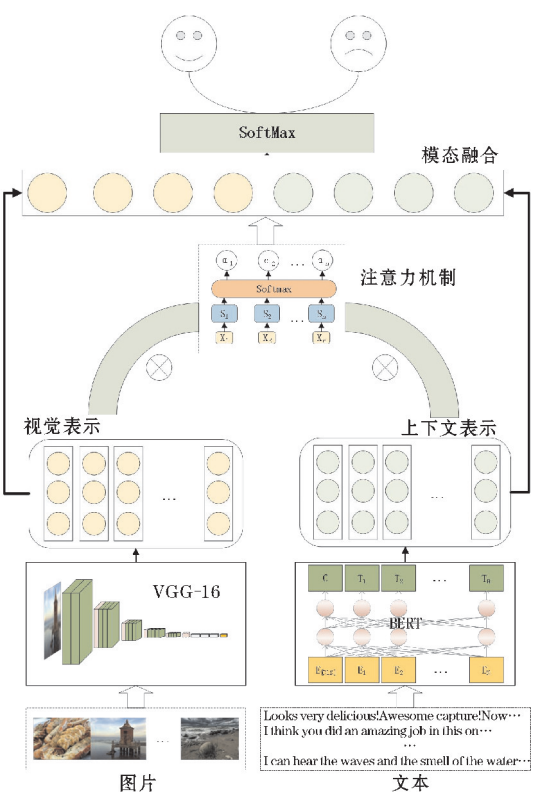


图 1 基于 BERT-VGG16 融合模型结构

1.1 文本处理模型

词嵌入(word embedding)是自然语言处理任务中的早期预训练技术,它将文本的单词或语句映射成向量数据,以便计算机识别和处理。2003 年 Bengio 等[13]提出神经网络语言模型(neural network language model,NNLM),研究者就不断提出新的词向量表示方法。2013 年 Mikolov 等[14]提出 word2vec 模型对词汇信息进行初始化,将各个词汇映射到相应的向量空间中,对细粒度的语义和句法都有较好的表达。2014 年 Pennington 等[15]提出 GloVe 模型,融合当时最新的全局矩阵分解法(LSA),做到全局词向量表达。以上的方法虽然可以将词和向量空间进行有效的对应,但是却无法解决一词多义等问题,如指代公司的“苹果”和水果中的“苹果”都会被经典词嵌入模型映射到同一个向量层,从而无法区分词语所代表的真正含义。

为了解决这个问题,Peters 等[16]提出了 ELMo(embedding from language model)模型,使用一种双向的 LSTM 语言模型(BiLSTM)在大型文本语料库上进行预训练,经过调整后的词向量更能表达词语在上下文中的具体含义。为了学习通用语言表示,Radford 等[17]提出了 GPT(generative pre-training model)模型,使用 Transformer 的 Decoder 替代 RNN。相比 ELMo,GPT 的特征抽取能力更强,但 GPT 采用的是单向语言模型,并没有把下文信息融合进来,导致部分信息缺失。随后,Devlin 等[18]提出的 BERT 模型结合 ELMo 和 GPT 的优点,不仅使用融合文本能力强大的 Transformer 作为特征抽取器,还使用双向语言模型,有效地提升文本语句的理解能力。BERT 模型结构如图 2 所示。

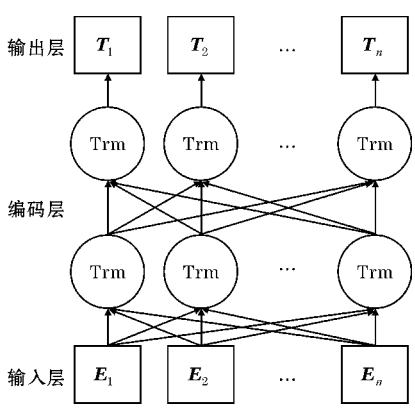


图 2 BERT 模型结构

图中,$E_1, E_2, \cdots, E_n$ 为文本数据的输入向量,Trm 为 Transformer 的 Encoder 神经网络,$T_1, T_2, \cdots, T_n$ 为数据处理后的输出向量。BERT 的预训练目的就是训练 Transformer 的 Encoder 网络,从而大幅提升准确率。通过两个无监督预测任务 Masked LM 和 Next Sentence

Prediction 预训练,微调预训练的 BERT 明显优于其他预训练的语言模型,在多个自然语言处理任务上取得了好的效果。

1.2　图像处理模型

图像处理是计算机视觉领域的基本任务之一,早期使用 KNN、SVM 等传统机器学习算法对图像进行分类,但并不能对图像的特征进行很好地提取。Krizhevsky 等[19]在 2012 年提出 AlexNet 网络,首次将深度学习用于大规模图像分类中,由此涌现一系列基于深度学习的图像分类模型。随着模型变得越来越深且结构设计越来越精妙,Top-5 的错误率也越来越低。牛津大学的 Simonyan 等[20]在 2014 年提出了 VGG 模型,根据模型层数的不同,又分为 VGG16 和 VGG19,该模型相比以往模型减少了卷积核尺寸,起到降参作用。同年,谷歌的 Szegedy 等[21]提出 GoogleNet 模型,该模型引入 inception 模块的新概念,将参数数量减少到 500 个。He 等[22]在 2015 年提出 RestNet 模型,引入残差思想,解决了网络在层数加深时优化训练的难题。

为有效地提取文本图片的视觉特征,本文采用从经典数据集 imageNet 上训练得到的 VGG16 模型作为图片预训练模型,通过迁移学习[23]将原始网络结构和参数应用到本数据集图片特征提取的任务中。在 VGG16 中,使用 3 个 3×3 卷积核来代替 7×7 卷积核,使用 2 个 3×3 卷积核代替 5×5 卷积核。主要目的是在保证具有相同感知野的条件下,提升网络的深度,在一定程度上提升神经网络的效果。VGG16 网络模型结构如图 3 所示。

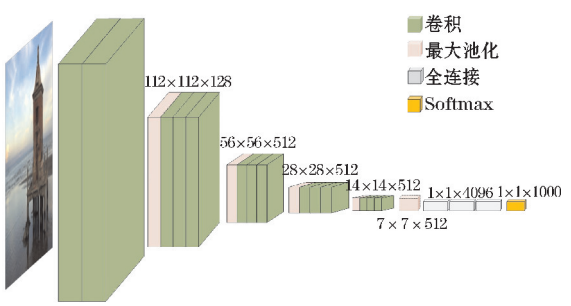


图 3　VGG16 网络模型结构

1.3　注意力机制

注意力机制[24]最早应用于图像处理领域,它借鉴了人类的注意力思维方式,通过快速扫描全局图像获取重点关注目标,从而投入更多注意力资源,以获取关注目标的细节信息。从本质上理解,Attention 是用权重代表目标信息重要程度,通过目标值 Value 与权重 $\boldsymbol{W}$ 进行加权求和的方式,计算注意力值。注意力机制模型如图 4 所示。

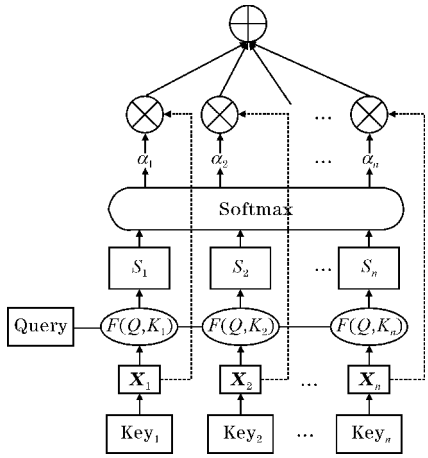


图 4　注意力机制模型

图中,$\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n$ 为输入向量,Query 为查询向量,用注意力变量 Key ∈ [1,N] 表示被选信息的索引位置,$F(Q,K_i)$ 为注意力打分函数,用来计算 Query 和 Key ∈ [1,N] 之间的相似性或相关性。$S_1,S_2\cdots,S_n$ 即为函数所求的函数值。

计算 F 函数通常有点积模型、Cosine 相似模型和 MLP 网络模型 3 种计算方式,计算公式为

点积模型:

$$F(Q,K_i)=\text{Query}\cdot\text{Key}_i$$

Cosine 相似模型:

$$F(Q,K_i)=\frac{\text{Query}\cdot\text{Key}_i}{\|\text{Query}\|\cdot\|\text{Key}_i\|}$$

MLP 网络模型:

$$F(Q,K_i)=\text{MLP}(\text{Query},\text{Key}_i)$$

$\alpha_1,\alpha_2,\cdots,\alpha_n$ 为所求得的 $S_1,S_2,\cdots,S_n$ 经过 SoftMax 函数进行数值转换求的值,这样不仅可以对数值进行归一化,而且可以更加突出重要元素的权重,计算公式为

$$\begin{aligned}\alpha_i&=p(z=i\mid x_i,\text{Query})\\&=\text{softmax}(S_i)\\&=\frac{\exp(f(x_i,\text{Query}))}{\sum_{j=1}^{N}\exp(f(x_j,\text{Query}))}\end{aligned}$$

求得的 α_i 也为 $\boldsymbol{X}_i$ 的值 Value_i 所对应的权重系数,⊗为权值相乘,⊕为求和,对所求得的 α_i 与 $\boldsymbol{X}_i$ 的 Value_i 进行加权求和,得到注意力数值:

$$\text{Attention}(\text{Query},\text{Source})=\sum_{i=1}^{L_x}\alpha_i\cdot\text{Value}_i$$

Attention 机制应用场景很广,Mnih 等[25]在 RNN 模型上使用 Attention 机制对图像进行分类。随后,Bahdanau 等[26]将 Attention 机制应用在机器翻译任务,取得很好的效果。之后,注意力机制越来越多地被应用到神经网络的各个领域,出现了很多注意力机制的变体。Yang 等[27]提出层级注意力网络模型,在文

本情感分类任务上取得明显的分类效果。Vaswani 等[28]提出自注意力机制(self-attention),将其应用到 Transformer 模型中,使分类效率大幅提升。Fu 等[29]提出多注意力神经网络,通过使用多注意力块发现不同模态之间的交互。

1.4 多模态融合

提取不同模态特征后,如何合并各个模态数据也是建模时需要解决的问题。多模态融合也是多模态研究中非常关键的研究点,它将抽取自不同模态的信息整合成一个稳定的多模态表征,表征信息的优劣往往决定最终融合模型的效果。常用的特征融合方法主要有 3 种:基于简单拼接的特征融合方法、基于张量的特征融合方法、基于注意力机制的融合方法。

简单拼接的融合是将不同模态数据统一在同一个向量空间维度,通过向量拼接的方式将所有模态数据进行融合。这种简单的拼接操作虽然能减少很多参数和计算量,但很容易造成各模块特征之间的信息丢失。张量是多维数组的泛概念,可看成是向量或矩阵的高阶扩展[30]。相对于简单向量求和的特征融合方法,通过对不同模态的数据进行张量积运算,能够充分保存模态间的交互信息。基于注意力机制的特征融合是在简单拼接融合方法前引入注意力机制,以达到充分利用模态间信息增益的目的,从而弥补简单拼接融合方法的不足。

本文采用的是基于注意力机制和张量计算的融合方式,特征融合前引入注意力机制,充分提取文本和图片的重要特征,再对各特征数据进行张量积运算,求得融合向量。特征融合模型结构如图 5 所示。

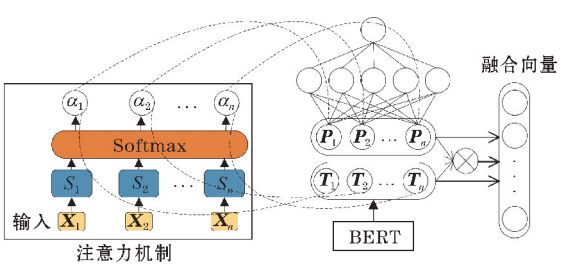


图 5 特征融合模型结构

在注意力机制模块中,使用点积模型计算出注意力权重值 $\alpha_1,\alpha_2,\cdots,\alpha_n$,再分别与图像经过神经网络模型 VGG16 所求的 $\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_n$ 和文本经过 BERT 所求的 $\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_n$ 进行向量运算,得到两个模态向量注意力值,通过张量积运算,求得融合向量。

图像的输出向量与权重 α_i 的计算公式:

$$\boldsymbol{P}_n=\boldsymbol{W}_p^{\mathrm{T}}\cdot\alpha_i$$

文本的输出向量与权重 α_i 的计算公式:

$$\boldsymbol{T}_n=\boldsymbol{W}_T^{\mathrm{T}}\cdot\alpha_i$$

特征融合公式:

$$\boldsymbol{M}=\boldsymbol{P}_n\otimes\boldsymbol{T}_n$$

2 实验与分析结果

2.1 实验数据集

实验使用的数据集来自 Flicker 网站中获取到的图片和对应的评论数据。为提高数据的多样性,分别在影视、自然、食品、动物、运动等不同类别下获取相关的图片和评论。通过对采集的数据集进行预处理和标注,分为积极和消极两种类别的情感表达。其中,消极图文数据共 4147 条,积极图文数据共 4596 条。为提高模型泛化能力,将整个数据集的 80% 用作模型的训练集,数据集的 20% 作为测试集。数据集分布如图 6 所示,数据集示例图如图 7 所示。

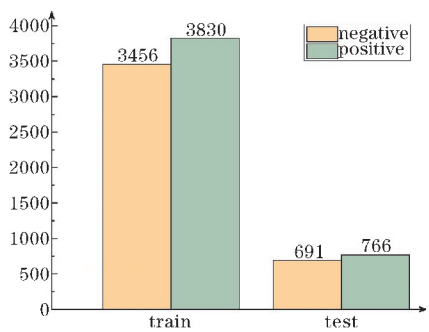


图 6 数据集分布图

图 7 数据集示例图

2.2 实验环境与参数设置

实验的硬件环境:CPU 为 i7-11800H,GPU 为 NVIDIA Geforce RTX3060,内存为 DDR4 16GB,CUDA 版本为11.3。使用的开发语言为 Python,开发工具为 Pycharm,开发环境为 Pytorch1.10.0+cu113。

实验使用的文本数据集为英文,故采用 Google 发布的英文版“BERT-Base,Uncased”为文本的预训练模型,Uncased 意味着文本在 WordPiece 标记化之前已被小写。此预训练模型有 12 层 Transformer,隐藏层维度为 768 维,Multi-Head-Attention 为 12,总计模型参数大小为110 M,在此预训练模型上进行微调。实验所设置的参数环境如表 1 所示。

表 1　实验参数环境

序号	参数	参数值
1	Batch_size	16
2	学习率	2e-5
3	优化器	Adam
4	最大序列长度	128
5	Dropout	0.2
6	epochs	10

2.3　实验评价标准

为验证本文模型在多模态情感分析中的有效性,使用精确率(precision)、召回率(recall)和 F1 值作为实验模型的评判标准,其计算公式分别如式(9)~(11)所示。其中,真正类(true postive,TP)为积极情感被预测为积极情感,假负类(false negative,FN)为积极情感被预测为消极情感,假正类(false postive,FP)为消极情感被预测为积极情感,真负类(true negative,TN)为消极情感被预测为消极情感。

精确率:

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}}$$

召回率:

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}}$$

F1 值:

$$\text{F1}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

2.4　对比实验

为验证实验模型的有效性,在相同的实验数据和环境下,对比其他单一模型和融合模型在本数据集中的表现,所参与对比的文本分类和图片分类模型如下。

文本分类模型:

(1)Text-CNN。网络结构简单,训练参数少,效率高,是文本分类任务中最常用的一种模型。

(2)BiLSTM。BiLSTM 是双向 LSTM 的组合,可以更好地捕获双向的语义依赖。

(3)BERT。使用 Transformer 代替传统 RNN 和 LSTM,有更深的层数和更好的并行性。

图像分类模型:

(1)CNN。一种前馈神经网络模型,图片数据经过一系列卷积层、池化层和全连接层后,得到分类概率。

(2)VGG16。缩小卷积核,通过不断加深网络结构来提升图像分类性能。

(3)RestNet。一种残差网络,经过堆叠可以构成一个更深的网络。

2.5　实验结果分析

实验首先测试各单模态模型分别在文本和图片上的情感分类效果,计算在测试集中的准确率、召回率和 F1 值,结果如表 2 所示。其中,对文本分类模型使用文本数据,对图像分类模型使用图像数据。

表 2　单模态模型评估结果对比

序号	模型	Precision	Recall	F1
1	TextCNN	0.6258	0.5530	0.5872
2	BiLSTM	0.6044	0.6318	0.6178
3	BERT	0.6970	0.6847	0.6908
4	CNN	0.5616	0.5237	0.5420
5	VGG16	0.6216	0.5832	0.6018
6	RestNet	0.6156	0.5654	0.5894

根据表 2,对于文本分类模型选择 BERT 模型,对于图像分类模型使用 VGG16 模型,分别使用 3 种融合方法对这两种模型处理后的数据进行融合,计算融合后的模型在测试集中的准确率、召回率和 F1 值,融合模型结果对比如表 3 所示。其中,SF 表示简单融合方法,TF 表示基于张量的融合方法,AF 表示基于注意力机制的融合方法。

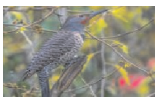
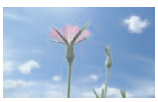
表 3　多模态融合模型评估结果对比

序号	模型	Precision	Recall	F1
1	BiLSTM-CNN-AF	0.6752	0.6054	0.6384
2	TextCNN-CNN-AF	0.6586	0.6256	0.6417
3	BERT-CNN-AF	0.6945	0.6385	0.6653
4	BERT-VGG16-SF	0.6543	0.5854	0.6179
5	BERT-VGG16-TF	0.6849	0.6089	0.6447
6	BERT-VGG16-AF	0.7259	0.6958	0.7105

通过实验结果发现,基于注意力机制的特征融合方法比另两种融合方法效果更好。为了增加此模型的对比程度,在实验中又随机组合了 3 组融合模型(序号 1,2,3),同时使用注意力机制融合方法进行模型融合。实验结果表明,基于 BERT-VGG16 加入注意力机制融合的模型结果更好。

为验证此模型在数据集中情感分类的效果,随机在测试集中挑选几组预测数据,部分样本抽取结果如表 4 所示。可以发现,当文本情感标注状态与图片情感标注状态相同时,模型所预测的结果与文本、图片情感标注的结果相同,预测目标情感的准确率更高。当文本情感标注状态与图片情感标注状态表达的情感相反时,模型将更偏重于文本所表达的情感状态。

表 4 样本抽取结果

文本数据	图片数据	文本情感标注	图片情感标注	整体情感标注	模型预测结果
If one day I can see him, I hope he is not happy, at least not happier than me.		negative	negative	negative	Negative(√)
The early bird gets the worm. I will also get up early every day and be as diligent as a bird.		positive	positive	positive	Positive(√)
The sun is so blinding, it's funny that the flower in the greenhouse never feels it, it never lacks pampering.		negative	positive	negative	Negative(√)
The happiest thing is to watch the sunset with my family.		positive	negative	positive	Positive(√)
To be merciful to your enemies is to be cruel to yourself, and to remain bestial is to be invincible in the law of the forest.		negative	positive	positive	Negative(×)

3 结束语

依托深度学习和各种神经网络模型技术,提出一种基于 BERT-VGG16 的多模态情感分析方法。该方法针对文本模态信息,使用 BRET 进行预训练,以获取包含上下文的语义信息。针对图片模态数据,使用 VGG16 神经网络对图片进行特征提取,然后对两种网络模型所提取的特征进行融合,通过在特征融合过程中加入注意力机制,能够有效地提高模型的重要特征指数,从而提高整个模型的准确率。由于实验设备和训练时间原因,本次实验并没有将所选择的分类模型和融合方法进行全局对比,下一步工作将进行更多的对比实验,同时也会选取其他数据集来测试本实验模型的泛化能力。

参考文献:

[1] Soleymani M, Garcia D, Jou B, et al. A survey of multimodal sentiment analysis[J]. Image and Vision Computing, 2017, 65: 3-14.

[2] Rao Y, Lei J, Wenyin L, et al. Building emotional dictionary for sentiment analysis of online news [J]. World Wide Web, 2014, 17(4): 723-42.

[3] Zhang L J, Wang S, LIU B. Deep learning for sentiment analysis: A survey[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018, 8.

[4] 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.

[5] Zhou C, Sun C, Liu Z, et al. A C-LSTM neural network for text classification[J]. arXiv preprint arXiv: 1511.08630, 2015.

[6] Ahmed K, Keskar N S, Socher R. Weighted transformer network for machine translation[J]. arXiv preprint arXiv: 1711.02132, 2017.

[7] Yang K, Xu H, Gao K. CM-BERT: Cross-Modal BERT for Text-Audio Sentiment Analysis[C]. Proceedings of the 28th ACM International Conference on Multimedia. 2020: 521-528.

[8] 陈鹏, 李擎, 张德政, 等. 多模态学习方法综述[J]. 工程科学学报, 2020, 42(5): 557-569.

[9] Peng Z J, Lu Y, Pan S, et al. Efficient Speech Emotion Recognition Using Multi-Scale CNN and Attention[J]. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 2021: 3020-3024.

[10] Lopes V, Gaspar A, Alexandre L A, et al. An AutoML-based Approach to Multimodal Image Sentiment Analysis[J]. 2021 International Joint Conference on Neural Networks(IJCNN), 2021: 1-9.

[11] Chen M, Li X. SWAFN: Sentimental Words Aware Fusion Network for Multimodal Sentiment Analysis[C]. Proceedings of the 28th International Conference on Computational Linguistics. 2020: 1067-1077.

[12] 林敏鸿, 蒙祖强. 基于注意力神经网络的多模态情感分析[J]. 计算机科学, 2020, 47(S2): 508-514.

[13] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. The journal of machine learning research, 2003, 3: 1137-1155.

[14] Mikolov T, Chen K, Corrado G, et al. Efficient es-

timation of word representations in vector space [J]. arXiv preprint arXiv:1301.3781,2013.
[15] 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.
[16] Peters M E, Neumann M, Logan IV R L, et al. Knowledge enhanced contextual word representations[J]. arXiv preprint arXiv:1909.04164,2019.
[17] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. 2018.
[18] 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.
[19] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems,2012,25:1097-105.
[20] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. CoRR,2015, abs/1409.1556.
[21] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015:1-9.
[22] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR),2016:770-778.
[23] Zamir A R, Sax A, Shen W, et al. Taskonomy: Disentangling task transfer learning[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018:3712-3722.
[24] 任欢,王旭光. 注意力机制综述[J]. 计算机应用,2021,41:1-6.
[25] Mnih V, Heess N, Graves A. Recurrent models of visual attention[C]. Advances in neural information processing systems. 2014:2204-2212.
[26] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. CoRR, abs/1409.0473,2015.
[27] 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.
[28] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Advances in neural information processing systems. 2017:5998-6008.
[29] Fu J, Zheng H, Mei T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017:4438-4446.
[30] Lee N, Cichocki A. Fundamental tensor operations for large-scale data analysis using tensor network formats[J]. Multidimensional Systems and Signal Processing,2018,29(3):921-960.

Multi-modal Sentiment Analysis Model based on BERT-VGG16

CHEN Hongsong, AN Junxiu, TAO Quanhui, ZHOU Jun
(College of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: Aiming at the problems that traditional sentiment analysis methods using only text data cannot fully dig out sentiment information, and the amount of information contained in monomodal data is limited, and cannot reflect the true emotional state well, a multi-modal sentiment analysis that introduces an attention mechanism is proposed model. First, the model uses pre-training models BERT and VGG16 to extract features from text data and image data, respectively. Secondly, in order to increase the weight of important features of each modal, the attention mechanism is introduced during feature fusion, and the fusion model can greatly increase the amount of data information. Experiment results show that the use of a multi-modal feature fusion model based on the BERT-VGG16 that introduces the attention mechanism has a significant improvement in the effect of sentiment analysis compared to single-modal and other multi-modal feature fusion models.

Keywords: sentiment analysis; multi-modal; BERT-VGG16 model; attention mechanism