计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(10):255−262 [doi: 10.15888/j.cnki.csa.009614]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

利用插值优化特征的多模态情感分析①

唐业凯[1], 冯　广[2], 杨芳捷[1], 林浩泽[2]

[1](广东工业大学 计算机学院, 广州 510006)
[2](广东工业大学 自动化学院, 广州 510006)
通信作者: 冯　广, E-mail: von@gdut.edu.cn

摘　要: 目前, 在多模态情感分析任务上, 存在着单一模态特征提取不充分、数据融合方法缺乏稳定性的问题. 本文提出一种利用插值优化模态特征的方法, 用于解决这些问题. 首先利用插值优化 BERT 和 GRU 模型提取特征的方式, 并使用这两种模型挖掘文本、音频、视频的信息. 其次, 用改进的注意力机制融合文本、音频和视频信息, 从而更稳定地实现模态融合. 该方法在 MOSI 和 MOSEI 数据集上进行实验. 实验结果表明, 使用插值能够在优化模态特征的基础上, 提高对多模态情感分析任务的准确率, 该结果验证了插值的有效性.

关键词: 插值; 特征提取; 注意力机制; 模态融合; 情感分析

引用格式: 唐业凯,冯广,杨芳捷,林浩泽.利用插值优化特征的多模态情感分析.计算机系统应用,2024,33(10):255–262. http://www.c-s-a.org.cn/1003-3254/9614.html

Multimodal Sentiment Analysis Using Interpolation Optimization Features

TANG Ye-Kai[1], FENG Guang[2], YANG Fang-Jie[1], LIN Hao-Ze[2]

[1](School of Computer Science, Guangdong University of Technology, Guangzhou 510006, China)
[2](School of Automation, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Currently, in multimodal sentiment analysis tasks, there are problems such as insufficient single modal feature extraction and lack of stability in data fusion methods. This study proposes a method of optimizing modal features that uses interpolation to solve these problems. Firstly, the interpolation-optimized BERT and GRU models are applied to extract features, and both of the models are used to mine text, audio, and video information. Secondly, an improved attention mechanism is used to fuse text, audio, and video information, thus achieving modal fusion more stably. This method is tested on the MOSI and MOSEI datasets. The experimental results show that using interpolation can improve the accuracy of multi-modal sentiment analysis tasks based on optimizing modal features. This result verifies the effectiveness of interpolation.

Key words: interpolation; feature extraction; attention mechanism; modal fusion; sentiment analysis

1　引言

在互联网发展迅速的今天, 人们每天都会在社交平台上产生许多具有丰富情感信息的多模态数据. 这些数据通常是人们对某种事物的观点和评价, 由于具有数量级大、形式多样的特点, 商家和管理者需要通过使用人工智能技术才能快速分析用户对他们产品的评价, 从而了解产品在用户心中的喜爱程度, 以便做出下一步决策. 该分析功能的人工智能技术是情感分析. 而多模态情感分析是指该情感分析任务的数据来源是包括文本、音频、视频 3 个模态的数据. 通过分析这 3 种模态的数据完成情感分析. 目前, 该任务具有重要的价值和意义[1].

① 基金项目: 国家自然科学基金重点项目 (62237001); 广东省哲学社会科学青年项目 (GD23YJY08)
收稿时间: 2024-02-21; 修改时间: 2024-03-19; 采用时间: 2024-04-23; csa 在线出版时间: 2024-08-21
CNKI 网络首发时间: 2024-08-22

目前多模态情感分析任务中, 主要存在以下问题.

(1) 目前多模态情感分析任务的研究者们专注于多模态数据融合, 而忽略了对各个模态的特征提取的充分性[2]. 特征提取不充分会导致单一模态下部分信息不完整影响情感分析的准确率.

(2) 目前模态融合使用数据拼接、注意力机制、多层感知机等方法实现, 但是这些方法跨模态融合时, 存在异构数据融合不稳定, 生成的特征表达可能出现异常导致情感分析任务的准确率不高.

因此, 在前人的研究基础上, 我们提出了利用插值优化特征提取的方法并使用一种改进的注意力机制进行模态融合, 解决上述问题. 我们的工作如下.

(1) 我们在模态融合之前, 利用了插值优化了 BERT 和 GRU 模型, 使该模型提取的特征更具充分性. 同时, 插值是一个信息结合的过程, 插值后的特征具有更丰富的信息量.

(2) 利用一种改进的注意力机制进行模态融合. 这种改进的注意力机制属于一种晚期融合. 它能保留各自模态的情感信息, 同时又能实现模态间信息的交互. 从而更稳定地实现模态融合.

我们使用上述方法优化了特征提取并解决了模态融合不稳定的问题. 我们使用这种方法提高了情感分析任务的准确率. 实验部分证明了效果是显著的.

2 相关工作

多模态情感分析的研究工作涉及自然语言处理、语音提取、计算机视觉等多个交叉领域. 主要研究工作是特征提取和模态融合两个模块.

在特征提取方面, Poria 等[3]利用 CNN[4]提取各个模态的特征, 利用这些特征向量对多核分类器进行训练得到情感结果. 陈敏[5]分别进行特征提取设计不同权重, 通过引入加权矩阵对各模态进行有效提取. Chen 等[6]关注时间步长对语音特征提取的作用, 提出利用门控多模态嵌入模型过滤音频数据提取特征.

在模态融合方面, Zadeh 等[7]研究出一种在 3 种模态内使用多视角序列学习, 从而挖掘了模态内的前后文关联的信息, 并用神经网络进行多特征融合的记忆融合网络 (MFN). Zadeh 等[8]又用分层次动态融合模态的方法, 提出了 GMFN 模型, 在 MOSEI 数据集上实现情感分析任务, 取得了更好的效果. Hazarika 等[9]将各个模态的数据映射到模态共享空间中, 用这种方法实现模态间信息交互和模态融合. Zhang 等[10]提出一种层次和选择交互的注意力模型 (HISA), 通过使用多个注意函数联合关注不同模态的特征来实现模态融合. Wu 等[11]通过增加多头注意力机制头数量的变化和使用门控信息通道替代前馈层优化特征提取, 通过张量融合网络融合. Yang 等[12]提出了 CM-BERT 模型, 在文本和音频两个模态上进行注意力融合实验. 将注意力机制用到模态融合任务上, 把音频模态的特征映射到文本模态上, 实现了模态信息注入, 提高模态信息表达的丰富性, 从而增强了情感分析分类的准确性. Tsai 等[13]提出多模态 Transformer 模型, 使用模态间定向交叉注意力进行模态融合. Han 等[14]提出的注意力机制融合方法, 在模态融合前先用门控单元改善单模态的表示, 再使用注意力提取多模态之间的相互信息, 实现了效果更好的情感分析.

还有部分研究者使用其他方法进行融合, Sun 等[15]用多层感知机 MLP 进行 3 种模态的融合, 用感知机在模态内先进行序列融合再进行模态间融合, 情感分析准确度较其他模型略低. Majumder 等[16]通过设置多种不同的双模态组合融合特征, 再将不同的组合进行模态融合. Han 等[17]通过使用融合结果再次与单模态信息交互的方法, 把模型和多模态情感分析任务同时训练, 从而提高模型情感分析性能.

综上所述, 在特征提取方面, 研究者们通过改变特征提取器和提取权重从而优化提取的模态特征. 在模态融合方面, 研究者们通过设计模态间交互与融合方式, 提高了模型性能. 但是模态特征表达不充分和数据融合不稳定问题依然存在.

3 利用插值优化特征的多模态情感分析方法

基于之前研究者的工作和目前存在的问题, 本文利用插值优化了特征提取的过程并结合了一种改进的注意力机制实现多模态数据融合和情感分析的方法. 该模型框架图如图 1 所示, 主要分为数据处理、插值优化特征、模态融合、情感分析 4 个部分. 其中, 绿色框代表增强器, 蓝色框是 BERT、GRU 模型编码和提取时序信息. Mix layer 是插值模块的混合层.

(1) 数据处理. 对文本、音频、视频数据进行特征提取和处理, 目的在于尽可能挖掘数据的信息量.

(2) 插值优化特征. 利用提出的混合层对模态内信息进行插值, 用这种方法优化特征提取生成的隐藏特

征, 达到丰富特征信息量的目的.

(3) 模态融合. 使用一种改进的注意力机制对文本、音频、视频的隐藏特征进行晚期融合, 保证模态稳定性的同时能够使模态间的信息得以交互.

(4) 情感分析. 将注意力机制融合模态的输出特征利用分类器进行情感分类.

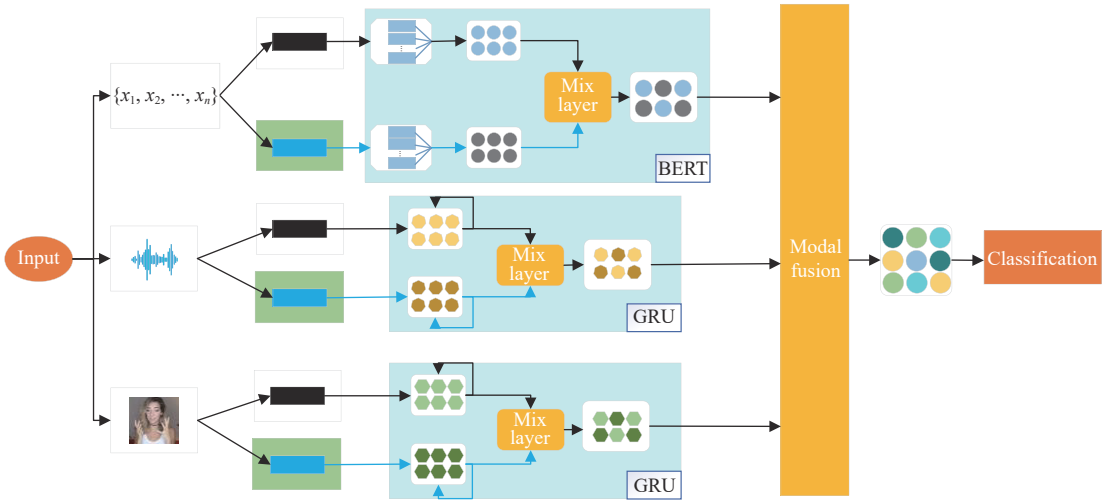

图 1 插值优化特征提取方法模型流程图

3.1 任务描述

给定的一组相关的多模态数据$S=\{X, A, V\}$, 包含一个句子$X \in R^{L_t}$, 一段音频$A \in R^{L_a \times d_a}$, 一个视频$V \in R^{L_v \times d_v}$, L_t、L_a、L_v分别表示文本、音频、视频的序列长度. d_a、d_v表示音频和视频的表示向量维数. 多模态情感分析任务的目标是根据给定的 S, 提取并整合 S 中不同模态的信息, 形成 S 的特征表示, 再利用得到的特征表示预测该组特征的情感标签 y.

3.2 数据处理

数据处理的任务是把 S 中的文本、音频、视频进行编码, 并在最后一层隐藏层输出h_t, h_a, h_v作为对应模态的特征表示. 除此之外, 为了利用插值挖掘更丰富的语义信息, 需要对原始数据做相应的处理.

3.2.1 文本数据的处理

首先将文本输入到文本增强器当中. 文本增强器是对输入的文本使用一种预设的文本数据增强方法的模块. 这些方法大多是常见的文本增强方法, 如同义词替换、随机插入、回译[18]等.

输入的文本并定义原始文本为 $X=\{x_0, x_1, x_2, \cdots, x_t\}$. L_t是文本序列长度. F定义为一种文本增强方法, 可以是同义词替换、回译、随机插入或是保持不变的任意一种. 则文本增强的方法可以表示为:

$$X' = \{x'_0, x'_1, x'_2, \cdots, x'_t\} = F \odot X \tag{1}$$

其中, $\odot$表示函数作用于集合中每一个元素.

构造文本模态数据集合$G=\{X, X'\}$. $G \in R^{2 \times L_t}$, 其中L_t为句子的序列长度.

将G双通道输入到 BERT 模型的嵌入层 (embedding), 对每一个单词生成相应的 token. 将生成的 token 输入到编码层 (encoder) 生成的文本特征用$T=\{h_t, h'_t\}$表示, 其中$T \in R^{2 \times L_t \times d_t}$. 嵌入层中有位置编码模块, 因此完成后可得到具有时序信息的文本特征.

3.2.2 音频和视频的数据处理

与文本类似, 将音频、视频数据输入到对应增强器. 在音频和视频增强方法上, 我们采用高斯噪声法, 即利用高斯噪声使声音和图像模糊化. 使用增强器的具体流程与文本类似. 得到的音频和视频信息为$A \in R^{2 \times L_a \times d_a}$、$V \in R^{2 \times L_v \times d_v}$.

音频和视频数据特征提取采用门控循环单元 (GRU)[19] 提取时间序列信息. 假设一段音频时间序列长度为 L_a, 每个时间步的隐藏特征长度为 d_a, 则该音频模态下的数据表示为$A \in R^{2 \times L_a \times d_a}$, 视频表示为$V \in R^{2 \times L_v \times d_v}$. 我们使用 GRU 捕捉音频和视频的时序特征:

$$\{h_a, h'_a\} = \mathrm{GRU}(A, \theta_a) \tag{2}$$

$$\{h_v, h'_v\} = \mathrm{GRU}(V, \theta_v) \tag{3}$$

其中, h_a 和 h_v 表示音频和视频的时序特征, A、V 是对

应模态的数据, θ_a、θ_v表示 GRU 模型的参数.

在数据处理阶段, 对每一种模态的数据都使用增强器丰富信息量, 这些增强器令一组数据的增加了一个维度, 从而扩充了信息量, 总体上不增加数据量的大小.

3.3 插值优化特征

插值优化特征是指把多组有丰富信息量的相似特征融合成一组, 实现模态内输出信息量最大化. 该插值优化过程在图 1 中的混合层 (Mix layer) 中进行. 其中, 深蓝色为原始特征, 浅蓝色为真实特征, 黄色为生成的特征, 绿色表示因插值而改变的特征.

文本特征的优化过程在 BERT 内进行. 音频、视频特征的优化过程在 GRU 进行提取时序信息后.

给定一组数据处理结束的单模态隐藏特征$\{h_m, h'_m\}$, $m \in \{t,a,v\}$. 设置保留权重参数$\lambda_{m\max}$和混合权重参数$\mu_m = \{\mu_{m1}, \mu_{m2}, \cdots, \mu_{m(n-1)}, \mu_{mn}\}$, λ_m参数由 *beta* 分布生成, 混合参数由迪利克雷分布生成, 这些参数用于插值的混合:

$$\lambda_m = beta(a,b) \tag{4}$$

$$\lambda_{m\max} = \max(\lambda_m, 1-\lambda_m) \tag{5}$$

$$\mu_{m1}, \mu_{m2}, \cdots, \mu_{mn} = \text{Dirichlet}(n) \tag{6}$$

其中, a, b, n 为超参数, 且为了简化该方法, 取 n=1. max 函数为取最大值函数, Dirichlet 为迪利克雷分布.

插值的目标是在保证原始特征不被破坏的情况下, 通过插入扰动特征丰富特征的信息量. 因而我们希望保留权重应比混合权重大, 故取$\lambda_{m\max}$保证其权重大于扰动特征. μ_m是扰动权重参数, 混合层的混合方法如下:

$$mix = \sum_{j=1}^{n} u_{mj} h'_{mj} \tag{7}$$

$$H_m = \lambda_{m\max} h_m + (1-\lambda_{m\max}) mix \tag{8}$$

其中, mix 为混合的隐藏特征, $\lambda_{m\max}$为保留权重系数.

$H_m \in R^{L_m \times d_m}$即混合层输出的隐藏特征. 单一模态特征经过提取时序信息和隐藏特征的插值, 该特征有更丰富的语义信息.

插值优化的在文本中的有效性解释为插值过程在学习原始数据附近的文本, 从而使混合文本相比于原始数据形成的文本更接近于真实文本[20].

在多模态特征表示中, 插值优化有效性解释如图 2 所示.

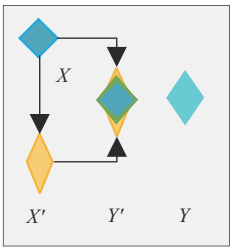

图 2 插值优化有效性解释

设原始特征为 X, 对应的真实特征为 Y. 由于特征提取技术的有限性, 原始特征 X 与 Y 相似而不完全一致. 通过增强器生成 X'并与 X 进行插值形成的 Y'. 从图 2 可以看出, 保留与 X 相似的情况下 Y'具有更丰富的信息量和与真实标签更为近似的特征表示.

3.4 模态融合

模态融合 (modal fusion) 是指将 3 种模态特征融合, 以获取更具有全面性和表征力的信息. 在本文中, 我们使用一种改进的注意力机制完成模态融合任务, 流程如图 3 所示.

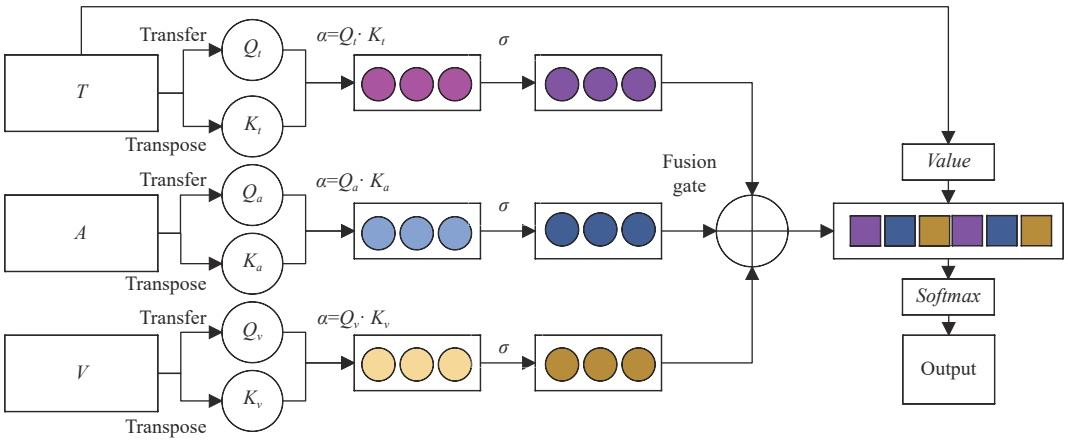

图 3 注意力机制模态融合模块

我们改进 Transformer 模型中的自注意力机制方法[21], 用输入的隐藏信息通过线性变换成 *Query* (Q)、*Key* (K)、*Value* (V). 通过计算 Q、K 之间的关联性得到注意力权重, 并将得到的权重作用于 *Value*, 从而实现模态融合, 并将结果作为模态融合的特征表示.

首先使用卷积神经网络[4]对齐各个模态的隐藏特征长度为 d, 保证注意力机制信息间交互时的稳定性.

$$Conv_m = Conv1D(d_m, d) \tag{9}$$

$$M = Conv_m(H_m) \tag{10}$$

其中, $Conv_m$ 是针对不同模态卷积网络参数的设定, $M \in \{T, A, V\}$为对齐后的隐藏特征, H_m为未对齐前的隐藏特征. $m \in \{t, a, v\}$ 为文本、音频、视频模态中的一种.

图 3 表示注意力机制模态融合的流程. 改进的自注意力机制的 Q、K 由各个模态的序列长度和隐藏特征信息构成, V 则由文本模态的序列长度和隐藏特征构成. 各个模态的序列长度和隐藏信息由式 (11) 生成相应的 Q_t、Q_a、Q_v, K_t、K_a、K_v 由式 (12) 构成. 在 Q_m、K_m 矩阵相乘时, 同模态的同构的信息进行交互, 保证注意力的稳定性. 同时, 用激活函数 ReLU 激活 α 得到注意力 β:

$$Q_m = \begin{cases} T, & m = t \\ A, & m = a \\ V, & m = v \end{cases} \tag{11}$$

$$K_m = \begin{cases} T^{\top}, & m = t \\ A^{\top}, & m = a \\ V^{\top}, & m = v \end{cases} \tag{12}$$

$$Value = T \tag{13}$$

$$\alpha_m = Q_m \circ K_m \tag{14}$$

$$\beta_m = \sigma(\alpha_m) \tag{15}$$

其中, $m \in \{t, a, v\}$, t、a、v 分别表示文本、音频、视频模态. $T^{\top}$、$A^{\top}$、$V^{\top}$ 表示 3 种模态特征的转置. $\circ$为矩阵间的内积. σ为激活函数.

从实验中我们发现各个模态所包含的信息重要性程度不同, 因此需要在融合门 (fusion gate) 融合时不能使用简单拼接, 而是需要在增加权重 w_m 用于调节 3 种模态注意力的权重. 权重 w_m 通过 *loss* 优化数值, 能训练出获取 3 种特征中最具表征力的信息. 最后再加上偏置权重矩阵 b. w_m 和 b 的值通过模型训练计算:

$$W_f = (w_t \circ \beta_t) \oplus (w_a \circ \beta_a) \oplus (w_v \circ \beta_v) \oplus b \tag{16}$$

其中, w_m, b 是由神经网络训练的参数矩阵. $\oplus$为矩阵对应位置数值相加.

按照式 (16) 训练并计算得到融合注意力权重 W_f 后, 对 W_f 和 *Value* 进行内积相乘生成融合信息 X:

$$X = W_f \circ Value \tag{17}$$

其中, $W_f \in R^{d \times L}$, $Value \in R^{L \times d}$, 所以形成的$X \in R^{d \times d}$. 其中 d 是数据对齐后隐藏特征的长度, L 是序列长度.

该模态融合过程中, 不同模态的异构特征只在融合门后进行晚期融合. 从而实现了该各模态特征更稳定地融合.

3.5 情感分析

得到生成的多模态信息, 在这步需要根据该信息和任务目标将该信息进行情感分类.

将多模态融合信息X输入到全连接层得到一个情感分数 Y_{pre}. Y_{pre} 用于分类. 我们采用回归的方式得到一个情感分数 Y_{pre}, 用这个情感分数 Y_{pre} 作为预测结果的标签.

$$Y_{\text{pre}} = Softmax(\text{FFN}(X)) \tag{18}$$

在计算损失时, 用真实标签 Y_{True} 和 Y_{pre} 采用 MAE 损失函数进行计算:

$$loss = \frac{1}{N} \sum_{i=1}^{N} \left| Y_{i\text{True}} - Y_{i\text{pre}} \right| \tag{19}$$

N 是一个批次的样本数, 计算出的损失利用反向传播优化模型中的参数. 得到的 Y_{pre} 标签用于后续分类. 若为二分类任务, $Y_{\text{True}} \in \{pos, neg\}$, 则将所有标签结果 Y_{pre} 映射为 pos、neg 两种. 若为七分类任务, $Y_{\text{True}} \in [-3, 3]$, 对应 7 种强度不同的情绪. 把标签结果 Y_{pre} 映射到[−3, 3]区间, 用分类器对 Y_{pre} 进行分类, 从而实现情感分析.

4 实验

在本节, 我们介绍实验相关的数据集、参数设置、评估指标、实验结果.

4.1 数据集

本文采用多模态情感分析数据集 MOSI 和 MOSEI 进行验证.

MOSI[22]是一个富含多模态信息的数据集, 旨在为情感分析、情感强度预测和主观性分析等领域提供强

有力的基准. 该数据集在 YouTube 上搜集获得, 涵盖了在线观点视频的多元素内容共 2 199 组, 包括文本、语音和视觉模态.

MOSEI[8]是一个用于社交互动情感分析的重要多模态数据集. 该数据集涵盖了文本、语音和视觉等多个感知模态, 标签包括情感类别和情感强度, 可用于多个任务, 如情感类别分类、情感强度预测以及跨模态情感分析. MOSEI 的规模较大, 包含 16 315 组训练数据、1871 组验证数据、4 654 组测试数据. 为该模型的训练和评估提供了丰富的数据. 相较于 MOSI, MOSEI 具有更大的数据量, 可以用于更多的任务.

4.2 参数设置

在数据处理阶段, 为了实验的简便性, 本实验在插值优化方法上对文本模态只使用同义词替换、音频和视频只使用了高斯噪声. 使用 Python 3.8 版本, 框架为 PyTorch 1.11.0, 显卡为 RTX4090. BERT 模型编码采用 BERT-base-uncased 预训练模型, 选取编码层第 12 层作为文本混合层. 在经门控单元后单独创建一层混合层作为音频和视频的混合. β 分布参数 a、b 使用 0.75. 文本模态隐藏特征长度为 768, 数据对齐时, 序列长度特征长度为 50, 隐藏特征长度为 30. 在注意力模块采用 dropout 为 0.25, 使用 Adam 优化器, 学习率设置为 2×10^{-5}.该模型在 30 个 epoch 中训练, Batch_size 设置为 32.

4.3 评估指标

评估指标我们选择二分类任务的准确率 ACC2、$F1$、ACC7 作为评估指标. ACC2 是指二分类任务的精确度、ACC7 是指七分类任务的精确度. $F1$ 分数是二分类任务下的一种分数, 计算公式如下:

$$F1 = 2\times(P\times R)/(P+R) \tag{20}$$

其中, P 是指精确率, 是二分类任务下模型分类正确的百分比, R 是指召回率, 是二分类任务下正例召回的百分比.

通过对比方法与基线模型的评估指标和对文本插值进行消融实验, 证明方法是有效的.

4.4 基线

(1) MFN[7]: 使用多视角序列学习, 从而挖掘了模态内的前后文关联的信息, 并用神经网络进行多特征融合的记忆融合网络.

(2) GMFN[8]: 分层次动态融合模态的方法, 该模型在 MOSEI 数据集上取得较好结果.

(3) MISA[9]: MISA 是一种学习特定模态的表示, 给予多模态数据一种观点帮助预测情感的模型

(4) Cube-MLP[15]: 采用 BERT-base 模型处理文本, 多层感知机进行模态融合的一种方法.

(5) CM-BERT[12]: 一种用注意力融合模态的方法, 该模型在 MOSI 数据集上的二分类取得了很高的分数.

(6) Self-MM[23]: 设计了一个标签生成模块获得单峰监督, 并引导子任务关注模态监督差异较大的样本的一个模型.

(7) CM-ERNIE [24]: 一种屏蔽多模态注意力动态调整文本和音频数据权重的方法, 让文本和音频模态的交互作用的微调预训练 ERNIE 模型.

4.5 实验结果和分析

为了验证插值优化和改进注意力融合模型的作用, 我们将我们的模型和基线模型进行对比, 结果如表 1.

表 1 不同模型在 MOSI 和 MOSEI 上的实验结果 (%)

Model	MOSI			MOSEI		
	ACC2	$F1$	ACC7	ACC2	$F1$	ACC7
MFN	77.4	77.3	34.1	76.0	76.0	—
GMFN	—	—	—	76.9	77.0	45.0
MISA	82.1	82.0	42.3	83.9	84.2	52.2
Cube-MLP*	83.6	<u>83.7</u>	44.5	<u>84.8</u>	84.8	51.5
CM-BERT*	83.0	82.4	42.4	73.5	74.1	51.9
Self-MM	83.0	82.7	<u>44.9</u>	82.7	<u>84.9</u>	**53.4**
CM-ERNIE	<u>83.9</u>	**84.0**	42.9	83.6	—	—
Ours	**83.9**	83.6	**45.1**	**85.9**	**85.4**	<u>52.7</u>

注: “*”表示相同情况下原论文方法复现的结果, 其他结果是从已发表论文中得到的, 其中最佳效果用粗体标出, 次佳效果用下划线标出

表 1 是该模型与其他模型对比的结果. 从表 1 可以看到, 我们的模型在 MOSI 数据集上 ACC2 和 ACC7 分别达到 83.9% 和 45.1%. 在 ACC2 分数上与 CM-ERNIE 模型相当, 比 Self-MM 模型高 0.9%. 在 ACC7 分数上较 Self-MM 模型提高 0.2%. 虽然提升较小, 但也能体现模型的优越性.

在 MOSEI 数据集上, 我们模型较于 Cube-MLP 模型在 ACC2、$F1$、ACC7 上提升 1.1%、0.6%、1.2%. 相较于 Cube-MLP, 我们的方法加入了插值优化了特征同时使用注意力机制代替了多层感知机, 提高的分数证明了方法的有效性. 相较于该模型而言, 在改进的注意力模态融合后, 我们加入了插值优化特征的方法, 实现了对多模态情感分析准确性的提升.

表 2 是该模型进行消融实验的结果. Base 是使用

串联方法拼接文本、音频和视频数据. 注意力融合则是第 3.4 节中提到的模态融合方法. 插值优化是第 3.3 节提到的特征优化方法.

表 2 模型在 MOSI 和 MOSEI 上的消融实验结果 (%)

方法	MOSI			MOSEI		
	ACC2	*F*1	ACC7	ACC2	*F*1	ACC7
Base	80.0	80.1	38.3	80.2	80.1	41.9
注意力融合	82.8	82.8	42.2	84.5	82.5	52.6
插值优化	80.8	80.6	39.0	80.8	80.7	42.0
注意力融合+插值优化	**83.9**	**83.6**	**45.1**	**85.9**	**85.4**	**52.7**

注: 最佳效果用粗体标出, 次佳效果用下划线标出

从结果可以看出, 改进的注意力融合方法相对于串联拼接方法提升很大, 在 MOSI 数据集上 ACC2、*F*1、ACC7 提升分别为 2.8%、2.7%、3.9%, 在 MOSEI 数据集上, 上述分数依次提升 4.3%、2.4%、10.7%. 改进的注意力融合方法在多模态情感分析上有一定成效.

而插值优化特征的方法在 Base 上使用插值方法时, 结果提升在 0.8% 左右. 可能的原因是拼接时是将数据串联, 并不进行模态交互, 由于信息量的增加, 分类器对大量信息的处理能力下降, 插值优化特征的效果体现不明显. 我们在注意力融合中加入插值优化时, 分类器处理的信息量减少, 插值优化特征的效果在 MOSI 的 ACC7 上提升 2.9%、在 MOSEI 的 ACC2 上提升 1.4%, 插值效果明显.

表 3 是检验插值优化方法在单一模态下作用的效果. 从单一模态的结果来看, 该方法在文本模态上两个数据集分别提升 0.4% 和 1.5%, 在视频模态上提升 0.8% 和 0.3%, 在音频模态上变化不大. 因此可以肯定该方法在文本、视频模态方面有正向作用. 插值优化音频的效果不如其他模态, 可能的原因是融合门中音频模态的权重小于文本和视频模态的权重, 导致优化音频的效果难以在该任务中体现.

表 3 优化方法进行消融实验的准确率 ACC2 (%)

方法	MOSI	MOSEI
不使用插值优化	82.8	84.5
插值优化文本模态	83.2	**86.0**
插值优化音频模态	82.8	84.3
插值优化视频模态	83.6	84.8
插值优化 (3种模态)	**83.9**	85.9

注: 当优化一种模态时, 其他两种模态不使用插值优化方法 (3种模态均使用插值优化方法除外)

值得一提的是, 改进的注意力机制方法是基于注意力机制改进的, 是一种能够稳定实现模态融合的方法. 而插值优化特征方法是基于提取并改变模态的原始特征而起到优化特征的效果, 因此插值优化特征的方法提升空间较融合方法小. 即便如此, 实验证明了插值优化特征的方法是有效的.

5 结论

在本研究中, 我们提出了一种利用插值优化特征的多模态情感分析方法, 并结合了一种改进的注意力机制形成一个情感分析模型. 通过用 MOSI 和 MOSEI 数据集对该模型进行实验验证, 插值优化特征丰富了特征表示的信息量, 而改进的注意力机制融合增加了融合过程的稳定性, 使得文本、音频和视频之间更充分地融合, 提高了模型对多模态信息的理解和利用能力. 实验结果表明, 我们的方法相较于基线模型表现更为出色. 特别是在丰富特征表示和提高准确率方面取得了显著提升, 可以证明插值优化特征的方法对多模态情感分析领域是有帮助的. 这一研究为多模态情感分析领域的方法创新和性能提升提供了实用的方法和实证支持. 未来的研究可以进一步优化插值, 比如优化增强器的增强方法或使用多种增强方法优化特征, 以适应更广泛的多模态任务场景.

参考文献

1 何俊, 刘跃, 何忠文. 多模态情感识别研究进展. 计算机应用研究, 2018, 35(11): 3201–3205.

2 Peng W, Hong XP, Zhao GY. Adaptive modality distillation for separable multimodal sentiment analysis. IEEE Intelligent Systems, 2021, 36(3): 82–89. [doi: 10.1109/MIS.2021.3057757]

3 Poria S, Cambria E, Gelbukh A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015. 2539–2544.

4 Cai GY, Xia BB. Convolutional neural networks for multimedia sentiment analysis. Proceedings of the 4th CCF International Conference on Natural Language Processing and Chinese Computing. Nanchang: Springer, 2015. 159–167.

5 陈敏. 融合文本和短视频的双模态情感分析 [硕士学位论文]. 南京: 南京邮电大学, 2020.

6 Chen MH, Wang S, Liang PP, *et al*. Multimodal sentiment analysis with word-level fusion and reinforcement learning. Proceedings of the 19th ACM International Conference on

Multimodal Interaction. Glasgow: ACM, 2017. 163–171.
7 Zadeh A, Liang PP, Mazumder N, *et al*. Memory fusion network for multi-view sequential learning. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018. 5634–5641.
8 Zadeh AAB, Liang PP, Poria S, *et al*. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne: ACL, 2018. 2236–2246.
9 Hazarika D, Zimmermann R, Poria S. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. Proceedings of the 28th ACM International Conference on Multimedia. Seattle: ACM, 2020. 1122–1131.
10 Zhang D, Wu LQ, Li SS, *et al*. Multi-modal language analysis with hierarchical interaction-level and selection-level attentions. Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME). Shanghai: IEEE, 2019. 724–729. [doi: 10.1109/ICME.2019.00130]
11 Wu J, Zhu TL, Zhu JH, *et al*. A optimized BERT for multimodal sentiment analysis. ACM Transactions on Multimedia Computing, Communications, and Applications, 2023, 19(2s): 91.
12 Yang KC, Xu H, Gao K. CM-BERT: Cross-modal BERT for text-audio sentiment analysis. Proceedings of the 28th ACM International Conference on Multimedia. Seattle: ACM, 2020. 521–528.
13 Tsai YHH, Bai SJ, Liang PP, *et al*. Multimodal transformer for unaligned multimodal language sequences. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 6558–6569.
14 Han W, Chen H, Gelbukh A, *et al*. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. Proceedings of the 2021 International Conference on Multimodal Interaction. Montréal: ACM, 2021. 6–15.
15 Sun H, Wang HY, Liu JQ, *et al*. CubeMLP: An MLP-based model for multimodal sentiment analysis and depression estimation. Proceedings of the 30th ACM International Conference on Multimedia. Lisboa: ACM Multimedia, 2022. 3722–3729.
16 Majumder N, Hazarika D, Gelbukh A, *et al*. Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowledge-based Systems, 2018, 161: 124–133. [doi: 10.1016/j.knosys.2018.07.041]
17 Han W, Chen H, Poria S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. ACL, 2021. 9180–9192.
18 Wei J, Zou K. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: ACL, 2019. 6382–6388.
19 Cho K, Van Merriënboer B, Gulcehre C, *et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha: ACL, 2014. 1724–1734.
20 Chen H, Han W, Yang DY, *et al*. DoubleMix: Simple interpolation-based data augmentation for text classification. Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju: International Committee on Computational Linguistics, 2022. 4622–4632.
21 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.
22 Zadeh A, Zellers R, Pincus E, *et al*. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv:1606.06259, 2016.
23 Yu WM, Xu H, Yuan ZQ, *et al*. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. Proceedings of the 35th AAAI Conference on Artificial Intelligence. AAAI, 2021. 10790–10797.
24 陶全桧, 安俊秀, 陈宏松. 基于跨模态融合 ERNIE 的多模态情感分析研究. 成都信息工程大学学报, 2022, 37(5): 501–507. [doi: 10.16836/j.cnki.jcuit.2022.05.003]

(校对责编: 孙君艳)