计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(5):195−202 [doi: 10.15888/j.cnki.csa.008475]

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

基于话语间时序多模态数据的情绪分析方法①

冯　广[1], 江家懿[1], 罗时强[2], 伍文燕[3]

[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)
[3](广东工业大学 网络信息与现代教育技术中心, 广州 510006)
通信作者: 伍文燕, E-mail: wuwy@gdut.edu.cn

摘　要: 长期以来, 传统的基于单模态数据情绪分析方法存在分析角度单一、分类准确率低下等问题, 时序多模态数据的分析方法为解决这些问题提供了可能. 本文基于话语间的时序多模态数据, 对现有的多模态情绪分析方法进行了改进, 使用双向门控循环网络 (Bi-GRU) 结合模态内和跨模态的上下文注意力机制进行情绪分析, 最后在 MOSI 和 MOSEI 数据集上进行验证. 实验表明, 利用话语间的时序多模态数据, 并且充分融合模态内以及跨模态上下文信息的方法, 能够从多模态特征和时序特征的角度进行情绪分析, 从而有效提高情绪分析任务的分类准确率.

关键词: 时序多模态数据; 双向门控循环网络; 注意力机制; 情绪分析

引用格式: 冯广,江家懿,罗时强,伍文燕.基于话语间时序多模态数据的情绪分析方法.计算机系统应用,2022,31(5):195–202. http://www.c-s-a.org.cn/1003-3254/8475.html

Sentiment Analysis Method Based on Temporal Multimodal Data Between Utterances

FENG Guang[1], JIANG Jia-Yi[1], LUO Shi-Qiang[2], WU Wen-Yan[3]

[1](School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China)
[2](School of Automation, Guangdong University of Technology, Guangzhou 510006, China)
[3](Center of Campus Network & Modern Educational Technology, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: The traditional sentiment analysis methods based on single-modal data have always had problems such as a single analysis angle and low classification accuracy. The analysis method based on temporal multimodal data provides the possibility to solve these problems. On the basis of the temporal multimodal data between utterances, this study improves the existing multimodal sentiment analysis method and uses the bidirectional gated recurrent unit (Bi-GRU) combined with the intra-modal and cross-modal context attention mechanism for sentiment analysis. The sentiment analysis is finally verified on the MOSI and MOSEI datasets. Experiments show that the method of using temporal multimodal data between utterances and fully integrating intra-modal and cross-modal context information can be applied to sentiment analysis from the perspective of multimodal and temporal features. By doing this, the classification accuracy of sentiment analysis can be effectively improved.

Key words: temporal multimodal data; bidirectional gated recurrent unit (Bi-GRU); attention mechanism; sentiment analysis

目前随着互联网的发展, 网络视频和社交平台日渐火热, 诸如国内的哔哩哔哩、抖音、快手, 国外的 YouTube、Twitter、TikTok 等, 这些平台上的网络视频包含了大量用户对某个事物所发表的观点和评价, 这些点评数据蕴含着用户的情绪信息, 挖掘这些用户的情绪信息不仅有利于平台商家对用户进行有针对性

① 基金项目: 国家自然科学基金 (71671048); 中国高校产学研创新基金 (2020ITA02013)
收稿时间: 2021-07-31; 修改时间: 2021-08-31; 采用时间: 2021-09-09; csa 在线出版时间: 2022-04-11

的产品推送, 对一些需要获得用户情绪状态的互联网服务型行业有所帮助, 而且在面对某些社会突发事件时, 也能够协助相关部门掌握社会舆论的走向[1].

目前大多数的情感分析方法存在以下两个问题: (1) 基于单模态数据的分析方法角度单一, 无法全面地反映人类复杂的情感表达. (2) 目前大部分基于多模态数据的情绪分析方法没有充分考虑并融合时序数据的模态内和跨模态的话语间时序信息关联, 导致情绪分析的准确率较低. 为解决以上问题, 本文使用时序多模态数据, 在前人的研究基础上, 改进了多模态情绪分析模型, 提出一种基于话语间时序多模态数据的情绪分析方法.

1 相关研究

情绪分析是一个涉及人工智能、计算机视觉、自然语言处理等多个多学科交叉的研究领域[2]. 早期的情绪分析大多是基于单一模态数据的, 目前主流方法是针对人脸表情和文本的情绪分析. Connie 等人[3] 使用由 3 个子网络构成的卷积神经网络在 CK+和 FER2013 数据集上进行人脸表情情感识别, 在 6 类的情感分类中得到良好的识别效果. 李婷婷等人[4] 针对微博短文本数据, 使用传统的 SVM 和 CRF 组合方法进行情感分析, 并选用不同的特征组合得到了最优的分析效果. 由于深度学习方法的兴起, 研究人员也开始在文本情感分析上使用深度学习模型. Chen 等人[5] 基于 TF-IDF 特征, 使用卷积神经网络进行文本情感分类, 相比传统的机器学习方法准确率得到了显著的提升. 曹宇等人[6] 使用 BGRU 对文本的上下文信息进行提取分析, 实验表明加入上下文信息后能够有效提高情绪分析的准确性.

由于单模态数据分析方法存在一定的局限性, 无法应对某些真实的场景, 因此研究者们开始探索使用两种或以上模态数据的研究方法, 同时因为网络视频这种多模态数据载体的兴起, 近年来针对视频数据集的多模态情感分析成为了一个研究热点.

在多模态情绪分析研究领域, 目前大多数研究是基于文本、语音和视觉 3 种模态信息. Baltrušaitis 等人[7] 将多模态机器学习分为模态表示、模态传译、模态对齐、模态融合和合作学习 5 个方面[8], 其中模态融合的效果对分析结果的准确性有很大的影响[9]. 模态融合主要分为特征级融合 (早期融合) 和决策级融合 (晚期融合), 二者的区别在于前者是将单模态特征直接进行融合后分析, 后者则是单独分析单模态特征后再对结果进行融合分析, 现在有研究者将这两种方式相结合并称之为混合融合. Pérez-Rosas 等人[10] 使用 OpenEAR 和 CERT 在 MOUD 多模态数据集中提取语音和面部的情感特征, 并且将单词与每个话语转录内的频率对应的值相关联, 得到加权特征图作为文本的情感特征, 最后将 3 种模态特征进行特征级融合后使用 SVM 分析, 在该数据集上得到良好的识别效果. Yu 等人[11] 针对中文微博数据, 使用 CNN 和 DNN 分别分析文本和视觉情感, 最后通过决策级融合的方法对分析结果进行融合, 在中文微博数据集上获得了最优的结果. Zadeh 等人[12] 提出一种张量融合方法 (TFN), 使用张量乘法将 3 种单模态特征数据融合在一起, 最后使用 MLP 神经网络进行分析预测, 在 MOSI 数据集上获得较好的准确率, 但这种方法没有利用话语级别的上下文时序信息特征, 而且时间复杂度和空间复杂度极高. 后来他们又提出了一种分层的动态融合图方法[13], 将三种模态信息两两组合, 首次在 MOSEI 数据集上获得较高的识别准确率且具有一定的可解释性. Poria 等人[14] 提出一种非端到端的方法, 使用两层 LSTM 网络分别对单模态和组合模态进行训练, 这种方法虽然利用了模态内的话语上下文信息, 但非端到端的方法增加了分析的复杂性, 不利于应用到实际.

近年来, 研究者们希望模型能够像人类在观察事物的时候, 能够把注意力集中在特征明显的部分, 因此原本被用于机器翻译领域的注意力机制被情绪分析领域的研究者们广泛关注, 并尝试在自己的模型中加入注意力机制, 使得模型能够关注数据中对情绪影响较大的特征. 朱烨等人[15] 融合了卷积神经网络和注意力机制对评论文本进行情绪分析, 实验表明使用注意力加权的方式识别准确率高于单一的 CNN 模型. Poria 等人[16] 对其原模型进行了改进, 先在单模态内部加入注意力机制, 随后在话语层面使用 LSTM 网络提取上下文信息, 然后再对拥有上下文信息的序列使用注意力模块, 相比他们原来的模型, 准确率有了较大的提升. Ghosal 等人[17] 提出了 MMMU-BA 模型, 对双模态使用注意力机制, 挖掘两个模态之间的上下文交互作用, 但该方法没有考虑增强单模态的上下文关联, 因此仍然存在改进的空间.

2 基于话语间时序多模态数据的情绪分析方法

时序多模态数据有两个特征: 一是每个数据样本

都存在 3 种模态可以分析, 二是句子上下文存在时序关联. 如图 1 所示, 以图中上文语句为例, 从文本模态“I did not like”来看, 该视频段说话者的情绪是消极的, 但结合语音模态和视觉模态分析可知文本模态信息相对冗余, 因此说话者所表达的真实情绪其实是积极的. 与单模态情感分析不同, 在同一个话语中, 不是所有模态都能发挥同等的作用, 所以多模态情感分析的难点在于如何有效整合不同模态的数据, 使模型既能发挥所有模态的作用, 也不会因为某个模态的冗余特征而影响到预测的结果. 同时, 以图 1 中的目标语句为例, 若只分析目标语句实际上难以准确判断此时说话者的情绪倾向, 但视频是由一系列的话语组成的, 每句话都具有特定的时间顺序, 与非时序数据不同, 视频中的每一段话语可能具有一定的关联性而且会彼此影响情感倾向[14]. 由此可见, 利用时序多模态数据进行情绪分析能够挖掘数据中不同模态、不同话语之间的内在关联. 因此本研究利用 MOSEI 和 MOSI 视频数据集的 3 种模态信息 (文本、声音、视觉) 进行多模态情绪分析研究, 并且利用卡内基梅隆大学开源的多模态数据处理 SDK, 在提取模态特征的同时保留视频上下文话语之间的时序特征, 通过注意力机制增强模态内和跨模态的上下文联系, 最后进行情绪分析, 形成一种基于话语间时序多模态数据的情绪分析模型.

图 1 时序多模态数据上下文影响情感分析的例子

本文基于话语间时序多模态数据的情绪分析模型框架如图 2 所示, 模型主要由以下 4 个部分组成.

(1) 单模态时序特征表示. 该部分主要是获取话语之间的上下文关联, 同时将各模态特征数统一到相同的维度.

(2) 模态内时序信息增强. 这部分的任务是增强上一层所得到的各模态内部的上下文信息关联.

(3) 双模态时序信息交互. 该层主要是对单模态时序信息特征进行跨模态融合, 不同模态两两组合, 并且挖掘跨模态的上下文关联.

(4) 情绪分类. 将各层输出的特征矩阵进行拼接后获得多模态融合信息, 进行情绪分类.

2.1 单模态时序特征表示

一个视频是由若干个视频片段组成的, 每个视频片段都具有时间顺序和特征. 在这一部分, 我们使用双向门控循环网络 (Bi-GRU) 来捕获视频片段的上下文语义信息. GRU 单元是 LSTM 单元的变体, 它将 LSTM 中的遗忘门和输入门合并成了一个更新门, 减少了参数的同时也能达到和 LSTM 相近的效果. 这里采用 Bi-GRU 则是为了更加充分地挖掘上文和下文对目标语句的影响. 假设一个视频有u个话语片段, 每个片段特征维度为d_m, 则某个模态下一个视频可以表示为$M \in \mathbb{R}^{u \times d_m}$, 其中$M, m \in \{T, A, V\}$分别为文本、语音和视觉模态. 以$u_t$表示视频中的一个话语片段, 设$x_t = [u_1, u_2, \cdots, u_t]$作为$Bi\text{-}GRU_m$的输入, 获得正向和反向输出序列的每个隐藏状态, 并将其拼接为一个隐藏状态h_t, 如下:

$$\overrightarrow{h_t} = \overrightarrow{GRU}(h_{t-1}, x_t) \tag{1}$$

$$\overleftarrow{h_t} = \overleftarrow{GRU}(h_{t-1}, x_t) \tag{2}$$

$$h_t = \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right] \tag{3}$$

经过 Bi-GRU 层的处理后, 可以获得一个隐藏状态序列$X = [h_1, h_2, \cdots, h_t]$, 该序列的每一个隐藏状态都包含受到上下文影响后该时刻话语的状态特征, 将其输入

到一个神经元数量为d的全连接层进行降维, 最后获得视频单模态时序特征$X_m = [x_{m1}, x_{m2}, \cdots, x_{mt}]$, $X_m \in \mathbb{R}^{u\times d}$.

2.2 模态内时序信息交互增强

这一部分主要是使用自注意力强化第 2.1 节中模态内部的上下文时序关联.

如图 3 所示, 首先将特征矩阵与其转置相乘, 获得上下文关联话语特征权重矩阵$S \in \mathbb{R}^{u\times u}$, 随后使用行 Softmax 对矩阵$S$进行归一化处理得到$S' \in \mathbb{R}^{u\times u}$, 即计算目标语句和其他语句的关联分数, 然后将该矩阵与原始特征矩阵进行矩阵相乘, 获得上下文联系的注意力表征矩阵$A \in \mathbb{R}^{u\times d}$, 强化与目标语句相关性高的语句特征的交互, 弱化相关性低的语句特征的交互, 最后将其与原始特征矩阵使用逐元素乘法, 得到模态内时序信息交互增强矩阵$O \in \mathbb{R}^{u\times d}$. 具体公式如下:

$$S = X \otimes X^{\mathrm{T}} \tag{4}$$

$$S'_{(i,j)} = \frac{\mathrm{e}^{S_{(i,j)}}}{\sum_{k=1}^{u} \mathrm{e}^{S_{(i,k)}}}, i, j = 1, 2, \cdots, u \tag{5}$$

$$A = S' \otimes X \tag{6}$$

$$O = A \odot X \tag{7}$$

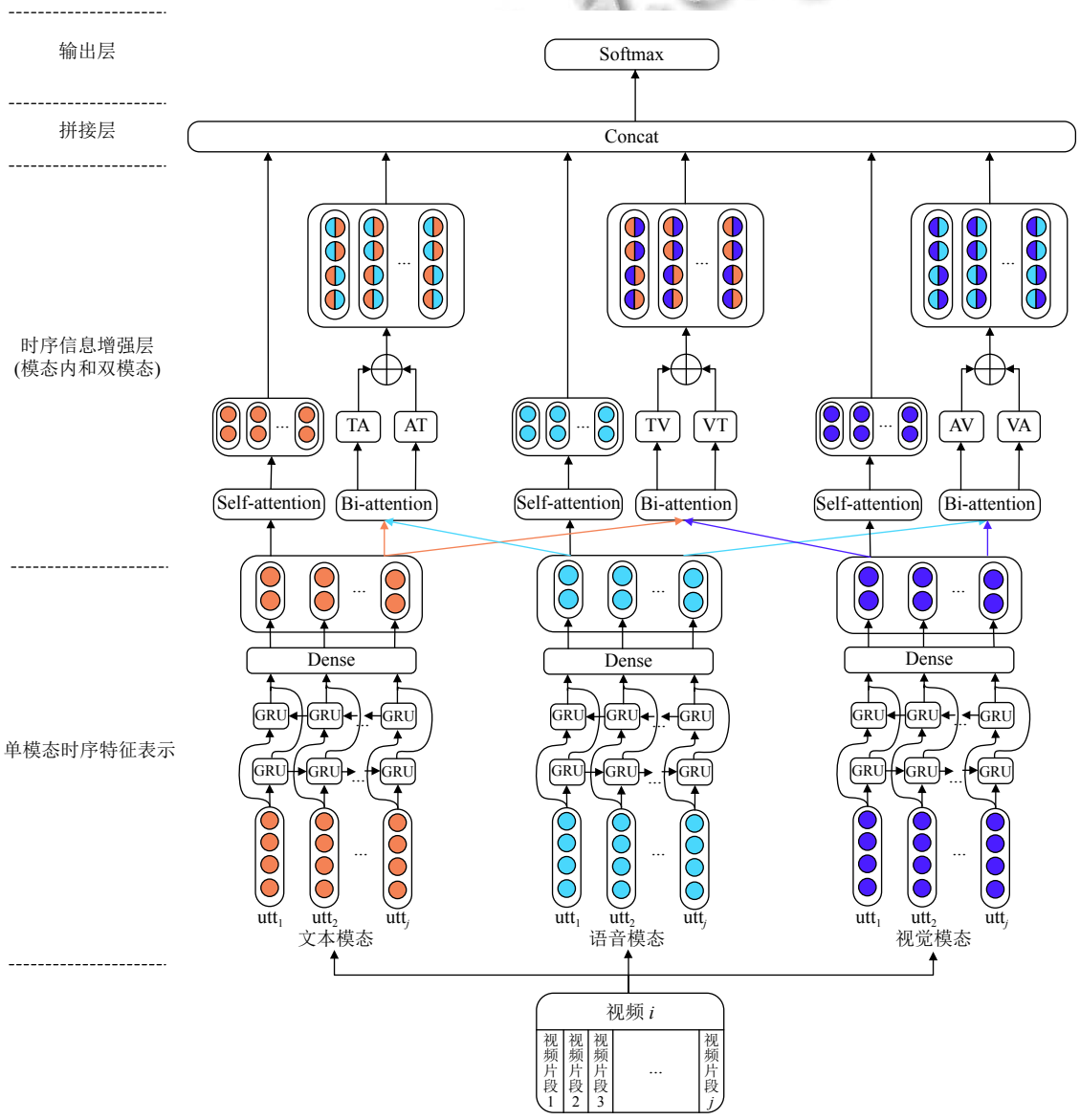

图 2 基于时序多模态数据的情绪分析模型框架图

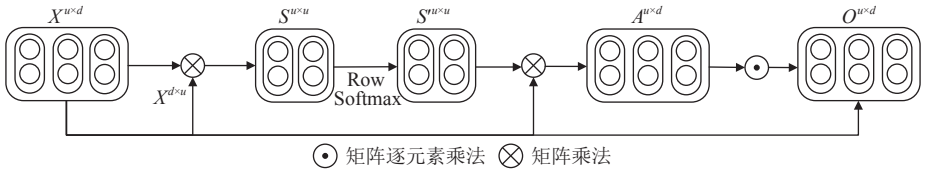

图 3 单模态内上下文交互增强方法

2.3 双模态时序信息交互

这一部分承接第 2.1 节, 挖掘双模态时序信息隐藏的内在关联, 分别计算两个模态相互之间的影响. 以文本和声音模态为例, 设一个视频内文本模态和声音模态的特征为$X_a, X_t \in \mathbb{R}^{u\times d}$, 首先分别计算各自模态每个话语对另一个模态每个话语的相关性分数矩阵$M_1, M_2 \in \mathbb{R}^{u\times u}$, 如$M_{1(1,2)}$代表声音模态第 2 句对文本模态的第 1 句的关联分数$(a \to t)$, $M_{2(1,2)}$则代表文本模态第 2 句对声音模态第 1 句的关联分数$(t \to a)$, 以此类推, 公式如下:

$$S_{at} = X_t \otimes X_a^{\mathrm{T}},\ S_{ta} = X_a \otimes X_t^{\mathrm{T}} \tag{8}$$

$$M_{1(i,j)} = \frac{\mathrm{e}^{S_{at(i,j)}}}{\sum_{k=1}^{u} \mathrm{e}^{S_{at(i,k)}}},\ i,j = 1,2,\cdots,u \tag{9}$$

$$M_{2(i,j)} = \frac{\mathrm{e}^{S_{ta(i,j)}}}{\sum_{k=1}^{u} \mathrm{e}^{S_{ta(i,k)}}},\ i,j = 1,2,\cdots,u \tag{10}$$

然后将相关性分数矩阵与关联模态特征矩阵相乘, 得到跨模态时序信息交互注意力表征矩阵$A_1, A_2 \in \mathbb{R}^{u\times d}$:

$$A_1 = M_1 \otimes X_a,\ A_2 = M_2 \otimes X_t \tag{11}$$

最后将其与目标模态的特征矩阵逐元素相乘得到$B_1, B_2 \in \mathbb{R}^{u\times d}$, 并拼接在一起, 作为文本-声音模态的双模态时序信息交互融合矩阵$O_{ta} \in \mathbb{R}^{u\times 2d}$:

$$B_1 = A_1 \odot X_t,\ B_2 = A_2 \odot X_a \tag{12}$$

$$O_{ta} = [B_1, B_2] \tag{13}$$

同理, 文本-视觉和声音-视觉的双模态时序信息交互矩阵也用同样的计算方法获得, 分别记为O_{tv}和O_{av}.

2.4 情绪分类

经过上述 3 层对 3 种模态的时序数据处理后, 我们可以得到 3 种模态内的时序信息交互增强矩阵和 3 个双模态时序信息交互矩阵, 将其进行拼接后得到最终的时序多模态融合特征矩阵$M \in \mathbb{R}^{u\times 9d}$并输入到全连接层进行降维, 得到$M \in \mathbb{R}^{u\times c}$, 由于本文为二分类, 因此$c$取 2.

记$x_{i,c}^{j}$为第i个视频中的第j个视频片段的第c个特征, 将其输入到 Softmax 分类器中, 公式如下:

$$P\left(\hat{y}_i^j = c\right) = \frac{\mathrm{e}^{x_{i,c}^{j}}}{\sum_{k=0}^{c} \mathrm{e}^{x_{i,k}^{j}}},\ i = 1,2,\cdots,V, j = 1,2,\cdots,u \tag{14}$$

$P\left(\hat{y}_i^j = c\right)$表示该视频片段为 0(消极) 和为 1(积极) 的概率, V和u分别代表视频样本总数和单个视频的片段总数, 即可完成情绪的分类. 训练过程还需要对分类的结果计算损失值以进行反向传播, 使用交叉熵函数进行计算, 公式如下:

$$loss = -\frac{1}{\sum_{i=1}^{V} U_i} \sum_{i=1}^{V} \sum_{j=1}^{U_i} \sum_{c=1}^{C} y_{i,c}^{j} \log_2\left(\hat{y}_{i,c}^{j}\right) \tag{15}$$

其中, U_i代表第i个视频中的有效视频片段数, 标签采用 one-hot 向量表示, 因此$y_{i,c}^{j}$代表第i个视频中第j个片段的真实标签, c代表类别, $\hat{y}_{i,c}^{j}$则是式 (14) 计算出的预测标签的概率值.

3 实验

3.1 数据集

本文使用卡内基梅隆大学的研究者提供的 MOSI 数据集和 MOSEI 数据集对模型进行验证.

(1) MOSI 数据集

该数据集包含了 YouTube 上的 93 个关于电影评论的视频, 每个视频都被切分为若干个视频片段并且打上情绪标签, 单个视频最多有 63 个话语片段, 总计 2 199 个片段. 本文使用 Poria 等人[14] 提供的处理过的 MOSI 数据集, 文本模态、声音模态和视觉模态的话语特征维度分别为 100、73 和 100, 由于只提供了训练集和测试集, 因此本文从训练集中抽取了一部分作为验证集, 得到训练集、验证集和测试集的视频数量为 52、10、31, 由于每个视频的话语片段数量不一, 因此对不足 63 片段数的样本进行补 0.

(2) MOSEI 数据集

该数据集包含了 3 228 个视频, 单个视频的话语片段数最多为 98, 总计 22 677 个视频片段. 本文使用卡内基梅隆大学提供的多模态数据 SDK 进行处理, 从原始数据集中提取包含话语间上下文时序信息的特征, 最终得到文本、声音和视觉 3 个模态的话语特征维度分别为 300、74 和 35, 训练集、验证集和测试集的视频数量为 2 250、300 和 678, 同样的, 对不足 98 个片段数的样本进行补 0.

3.2 参数设置

本实验在 Ubuntu 16.04 操作系统上完成, 内存大小为 32 GB, GPU 型号为 GTX3070, CPU 型号为 3.7 GHz 主频的 i5-9600k, 深度学习框架使用 TensorFlow 2.4 的 GPU 加速版本. 模型中提到的 Bi-GRU 网络的隐藏层单元数为 300, 后续全连接层神经元数量为 100, 训练批次 (batch_size) 大小为 64, 共迭代 (epoch) 50 次. 使用 Adam 优化器更新模型参数, 学习率为 0.001. 同时在训练过程中, 使用 dropout 降低过拟合, 对于 MOSI 和 MOSEI 数据集, 模型设置 dropout 值分别为 0.3 和 0.5. 最后使用准确率和 *F*1 值作为模型的评价指标.

3.3 基线模型选择

本文将选择以下模型作为本方法的基线模型进行比较.

(1) TFN[12]: 该模型由 Zadeh 等人在 2017 年提出, 直接将 3 种模态的特征数据统一到同一维度后, 进行张量乘法操作, 形成一个张量后输入到分析网络中, 没有考虑时序信息特征, 而且时间复杂度和空间复杂度都很高.

(2) MFN[18]: 该模型由 Zadeh 等人在 2018 年提出, 使用一种多视图顺序学习的神经网络结构, 使用 LSTM 网络挖掘一个话语中的前后文关联与跨模态交互.

(3) BC-LSTM[14]: 该方法是由 Poria 等人在 2017 年提出的一种非端到端学习方法, 使用双向 LSTM 网络先对单模态数据进行训练, 再将训练特征拼接起来作为多模态融合数据进行训练.

(4) GMFN[13]: 该模型由 Zadeh 等人在 2018 年提出, 它以分层的方式动态融合模态, 首次在 MOSEI 数据集上取得较好的结果.

(5) MMMU-BA[17]: 该模型由 Ghosal 等人在 2018 年提出, 它使用一种跨模态注意力机制充分融合了双模态的时序特征信息, 但并没有充分考虑到单模态内的时序信息的交互作用.

3.4 实验结果与分析

为验证多模态时序信息在情绪分析中的重要性, 本文首先是将模型中的每一个单一结构进行测试, 结果如表 1 所示.

表 1 模型中单一结构的效果对比 (%)

单一结构	MOSEI		MOSI	
	准确率	*F*1值	准确率	*F*1值
T	79.11	77.93	80.19	79.92
A	75.19	72.70	63.43	62.10
V	75.07	73.84	57.98	53.72
TT	79.40	78.80	80.72	79.84
AA	75.26	72.86	62.10	61.09
VV	75.27	73.84	61.17	54.92
TA	79.73	78.58	81.80	80.65
TV	79.45	78.20	80.05	79.79
VA	76.81	75.69	65.43	55.32
本文	**80.12**	**78.96**	**82.46**	**80.92**

从表 1 中单一结构 T、A、V 来看, 在 3 种模态中文本模态提供了最多的信息, 准确率和 *F*1 值都是最高的, 因此一般情况下通过文本模态可以大致确定说话者的情绪倾向. 当使用自注意力机制加强了模态内的上下文信息后 (表中单一结构 TT、AA、VV), 对 MOSEI 数据集而言, 3 种模态的准确率和 *F*1 值都有所提高, 但对于 MOSI 数据集而言, 声音模态存在较多的冗余数据, 加强了模态内上下文信息后准确率和 *F*1 值反而下降了. 单一结构 TA、TV、VA 则是双模态的上下文信息交互层, 可以看出跨模态的上下文信息交互能提供更好的识别效果, 但由于视觉模态和声音模态本身提供的信息较弱, 因此这二者的融合效果会比有文本模态的融合效果要差, 由此可见不是所有模态都能提供相同的分析效果, 甚至存在冗余的模态信息会对分析效果产生负面的影响. 最后则是将模态内的时序信息交互特征和双模态的时序信息交互特征拼接后进行分析, 得到的准确率和 *F*1 值都比前面所述的单一结构高.

本文模型与其他模型的对比如表 2 所示. 从表中可以看到, 前 3 个模型都没有利用到多模态数据的时序特征, 仅仅是针对单个话语进行训练与识别, 而 BC-LSTM 和 MMMU-BA 模型利用到了数据的时序特征, 准确率与 *F*1 值都有明显的提高, 证明了话语级的时序信息特征确实能够提高情绪分析的识别. 本文的方法对前人的模型进行了改进, 同时融合模态内的时序信

息特征和双模态的时序信息特征, 在 MOSEI 数据集上准确率比基准模型最高值提高了 0.32%, 而 *F*1 值提高了 1.96%, 在 MOSI 数据集上, 准确率提高了 0.15%, 而 *F*1 值虽然比最高值低, 但仍然比非话语级时序的分析高. 由此可见, 本文提出的方法是能够提高情绪分析的识别准确率, 同时模型的稳健性更高.

表 2 不同模型的效果对比 (%)

模型	MOSEI		MOSI	
	准确率	*F*1值	准确率	*F*1值
TFN	—	—	77.1	77.9
MFN	—	—	77.4	77.3
GMFN	76.9	77.0	—	—
BC-LSTM	77.64	—	80.3	—
MMMU-BA	79.80	71.37	82.31	**81.22**
本文	**80.12**	**78.96**	**82.46**	80.92

4 结语

本文提出了一种基于话语间时序多模态数据的情绪分析方法, 有效提取了模态内的时序信息交互特征和双模态的时序信息交互特征. 首先通过对模型中每一个结构进行单独的实验, 可以看出单模态数据提供的分析角度较为单一, 时序多模态数据的分析方法通过利用多模态特征和时序特征, 有效提高模型分析角度的全面性, 同时, 在加入了时序信息增强特征和双模态时序交互特征后, 更是明显提高了情绪分析任务的准确率. 最后与现有的模型进行比较, 证明了本文提出的方法在 MOSEI 和 MOSI 数据集上能够不仅有效提升了情绪分析任务的识别准确率, 还得到了更好的模型稳健性. 由此可见, 话语间的时序多模态数据蕴含了更多的情绪信息, 其特征的提取、模态的融合等会对识别效果产生显著的影响. 因此后续的工作将继续在多模态情绪分析这一领域, 在特征提取与模态融合的方向进行更深入的研究.

参考文献

1 王雨竹, 谢珺, 陈波, 等. 基于跨模态上下文感知注意力的多模态情感分析. 数据分析与知识发现, 2021, 5(4): 49–59.

2 刘继明, 张培翔, 刘颖, 等. 多模态的情感分析技术综述. 计算机科学与探索. 2021, 15(7): 1165–1182.

3 Connie T, Al-Shabi M, Cheah WP, *et al*. Facial expression recognition using a hybrid CNN-SIFT aggregator. Proceedings of the 11th International Workshop on Multi-disciplinary Trends in Artificial Intelligence. Gadong: Springer, 2017. 139–149.

4 李婷婷, 姬东鸿. 基于 SVM 和 CRF 多特征组合的微博情感分析. 计算机应用研究, 2015, 32(4): 978–981. [doi: 10.3969/j.issn.1001-3695.2015.04.004]

5 Chen JY, Yan SK, Wong KC. Verbal aggression detection on Twitter comments: Convolutional neural network for short-text sentiment analysis. Neural Computing and Applications, 2020, 32(15): 10809–10818. [doi: 10.1007/s00521-018-3442-0]

6 曹宇, 李天瑞, 贾真, 等. BGRU: 中文文本情感分析的新方法. 计算机科学与探索, 2019, 13(6): 973–981. [doi: 10.3778/j.issn.1673-9418.1806018]

7 Baltrušaitis T, Ahuja C, Morency LP. Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(2): 423–443. [doi: 10.1109/TPAMI.2018.2798607]

8 林子杰, 龙云飞, 杜嘉晨, 等. 一种基于多任务学习的多模态情感识别方法. 北京大学学报 (自然科学版), 2021, 57(1): 7–15.

9 Wu LZ, Oviatt SL, Cohen PR. Multimodal integration—A statistical view. IEEE Transactions on Multimedia, 1999, 1(4): 334–341. [doi: 10.1109/6046.807953]

10 Pérez-Rosas V, Mihalcea R, Morency LP. Utterance-level multimodal sentiment analysis. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Sofia: Association for Computational Linguistics, 2013. 973–982.

11 Yu YH, Lin HF, Meng JN, *et al*. Visual and textual sentiment analysis of a microblog using deep convolutional neural networks. Algorithms, 2016, 9(2): 41. [doi: 10.3390/a9020041]

12 Zadeh A, Chen MH, Poria S, *et al*. Tensor fusion network for multimodal sentiment analysis. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017. 1103–1114.

13 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: Association for Computational Linguistics, 2018. 2236–2246.

14 Poria S, Cambria E, Hazarika D, *et al*. Context-dependent

sentiment analysis in user-generated videos. Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 1: Long Papers). Vancouver: Association for Computational Linguistics, 2017. 873–883.

15 朱烨, 陈世平. 融合卷积神经网络和注意力的评论文本情感分析. 小型微型计算机系统, 2020, 41(3): 551–557. [doi: 10.3969/j.issn.1000-1220.2020.03.018]

16 Poria S, Cambria E, Hazarika D, *et al*. Multi-level multiple attentions for contextual multimodal sentiment analysis. Proceedings of 2017 IEEE International Conference on Data Mining. New Orleans: IEEE, 2017. 1033–1038.

17 Ghosal D, Akhtar S, Chauhan D, *et al*. Contextual inter-modal attention for multi-modal sentiment analysis. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 3454–3466.

18 Zadeh A, Liang PP, Poria S, *et al*. Multi-attention recurrent network for human communication comprehension. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018. 5642–5649.