文章编号：1673-3193（2024）03-0274-12

句子级时序卷积网络的多模态抑郁症识别方法

王烽飞[1]，卓广平[1]，周金保[1]，刘国强[1]，张光华[2]

（1. 太原师范学院 计算机科学与技术学院， 山西 晋中 030619；
2. 太原学院 智能与自动化系， 山西 太原 030032）

摘　要：针对多模态抑郁症模型在特征提取时，语句间关联性较弱，不同模态间的特征融合较为随意，在中文数据集上模型的泛化能力缺乏验证等问题，本文通过分析与抑郁症相关的音频、文本和视觉特征，提出了基于改进TCN模型的多模态抑郁症识别模型STCMN（Sentence-level Temporal Convolutional Memory Network），并将该模型应用于临床抑郁症辅助诊断当中。该模型首先使用残差块、GRU和Self-Attention的融合模块来提取不同模态下的句子级特征，增强了上下文联系，然后使用TCN模型来提取不同模态的全局特征，并使用Cross Attention对不同模态的全局特征以多模态融合特征为主进行融合，最后通过LogSoftmax层得到模型对抑郁症的识别结果。在DAIC-WOZ公开数据集上，本文所提出的方法对抑郁症识别的准确率达到了91.3%，精确率达到了93.6%，召回率达到了89.7%，其相关指标均优于其他方法，可以更好地满足临床医学的需求。在私有中文数据集MMD2022上，STCMN模型的识别结果仍为最优，表明该模型在中文抑郁症识别任务上具较好的泛化能力。

关键词：抑郁症；时序卷积网络；门控循环单元；自注意力机制；交叉注意力机制
中图分类号：TP391　　**文献标识码**：A　　**doi:** 10.3969/j.issn.1673-3193.2024.03.004
引用格式：王烽飞，卓广平，周金保，等. 句子级时序卷积网络的多模态抑郁症识别方法[J]. 中北大学学报（自然科学版），2024，45(3)：274-285.
WANG Fengfei，ZHUO Guangping，ZHOU Jinbao，et al. Sentence-level temporal convolutional networks for multimodal depression recognition[J]. Journal of North University of China(Natural Science Edition)，2024，45(3)：274-285.

Sentence-Level Temporal Convolutional Networks for Multimodal Depression Recognition

WANG Fengfei[1]，ZHUO Guangping[1]，ZHOU Jinbao[1]，LIU Guoqiang[1]，ZHANG Guanghua[2]

（1. College of Computer Science and Technology，Taiyuan Normal University，Jinzhong 030619，China；
2. Department of Intelligence and Automation，Taiyuan University，Taiyuan 030032，China）

Abstract：In the feature extraction of multimodal depression models，there are problems such as weak correlation between sentences，random feature fusion between different modalities，and lack of verification of the generalization ability of the model on the Chinese data set. By analyzing audio，text and visual features related to depression，this paper proposed a multi-modal depression recognition model STCMN（Sentence-level Temporal Convolutional Memory Network）based on improved TCN model. And the

收稿日期：2023-09-27
基金项目：山西省自然科学基金面上项目（201801D121147）；山西省重点研发计划项目（202202150401019）；太原师范学院研究生教育创新资助项目（SYYJSYC-2399）
作者简介：王烽飞（1998—），男，硕士生，主要从事自然语言处理方面的研究。
通信作者：卓广平（1972—），男，副教授，博士，主要从事大数据分析与挖掘、认知与智能等研究。E-mail：1643395248@qq.com。

model was applied to the auxiliary diagnosis of clinical depression. Firstly, the fusion module of residual block, GRU and Self-Attention was used to extract the sentence-level features under different modalities, which enhances the context connection. Then, the TCN model was used to extract the global features of different modalities. Cross Attention was used to fuse the global features of different modalities mainly with multi-modal fusion features. Finally, the recognition results of the model for depression were obtained through the LogSoftmax layer. On the DAIC-WOZ public dataset, the accuracy rate, precision rate and recall rate of the proposed method for depression recognition reach 91.3%, 93.6% and 89.7%, respectively. The related indicators are better than other methods, which can better meet the needs of clinical medicine. On the private Chinese dataset MMD2022, the recognition results of STCMN model are still the best, indicating that the model has good generalization ability in Chinese depression recognition tasks.

Key words: depression; TCN; GRU; self-attention; cross attention

0 引　言

抑郁症作为一种常见且严重的心理疾病，对患者的生活质量和社会经济的发展造成了严重影响。经研究发现，全世界有4.4%的人口患有不同程度的抑郁症。更糟糕的是，在全球新冠疫情大流行期间，全球精神障碍类疾病的负担更加沉重，重度抑郁症和焦虑症的病例分别增加了27.6%和25.6%，抑郁症患者激增约5 300万人[1]。因此，提高抑郁症早期识别的准确率具有重要的临床意义。

目前，抑郁症诊断主要依赖于临床医生的主观评估和患者的自述信息，这种方法存在着主观性高、诊断时间长、容易受到个体差异影响等问题。近年来，随着信息技术和机器学习的快速发展，深度学习在计算机视觉、语音识别和自然语言处理上取得了显著成效，通过将多模态数据（如语音、图像、文本流信息）与深度学习算法相结合来提高抑郁症识别的准确率的方法已成为一种值得探索的研究方向。

近年来，对抑郁症识别模型的研究无论在单模态还是多模态，研究人员们都取得了很大的进展。在使用音频数据作为模型的输入中，McGinnis等[2]提出了一种使用3 min语音数据来识别抑郁症的方法；Di Matteo等[3]通过手机程序收集患者音频数据；Flores等[4]应用语音迁移学习模型对语音片段分类后再进行抑郁症识别。文本作为识别抑郁症的主要数据之一，Tlachac等[5]以短信和推文作为数据源来预测参与者的健康问卷分数，Senn等[6]比较了3种BERT变体和4种BERT变体组合对12个临床访谈问题的应答转录文本，并根据其结果进行抑郁症分类。在根据视觉特征对抑郁症识别的研究中，Jazaery等[7]提出了利用三维卷积神经网络自动学习人脸区域的两个不同尺度的时空特征对人脸表情的局部和全局时空信息进行建模，以预测抑郁程度；Wang等[8]通过视频提取关键面部特征来探索抑郁症患者与正常人在相同情境下面部表情的差异。这些单模态方法具有参数量小、维度少和模型构建简单的特点。语音数据作为最重要的数据源之一，可以使分类模型同时利用音频特征和文本特征。Asgari等[9]通过对文本内容和声音韵律的研究，使抑郁症的临床筛查能力得到了提升；Rodrigues等[10]提出了一种用于抑郁症检测的语音和语言表征的多模态融合方法，其中使用CNN和LSTM相结合的方法估计PHQ分数；Toto等[11]提出音频辅助BERT（AudiBERT）证明了使用语音技术从非正式对话中筛查抑郁症的可行性。但由于个体的多样性和复杂性，人们并没有一个统一的标准来表达自己的情绪。比如有些人喜欢用语言来表达自己的观点，有些人则不善言辞，在这种情况下，不仅需要根据他们不多的言语得到文本和声学特征，还需要通过他们的面部表情来判断他们的情绪和状态，这种多模态的情感分析相比单模态方法实现了信息互补，提高了模型的鲁棒性和识别精度。Haque等[12]将时域卷积网络（Temporal Convolutional Network，TCN）模型用在多模态抑郁症识别任务中并得到较好的实验结果；Ray等[13]提出了基于多级注意力的多模态抑郁症预测网络，在学习模态内和模态间相关性的同时，融合了来自音频、视频和文本模态的特征；Cao等[14]将GRU与Bimodal Attention相结合提取上、下文的多模态特征信息。除了常用的卷积神经网络和循环神经网络，

为了解决长期依赖问题而设计的长短期记忆网络(Long Short Term Memory, LSTM)也在多模态抑郁症识别任务中表现出了强大的能力,其中,Flores等[15]将多个预训练迁移学习模型和带有自注意力的双向LSTM相结合,显著提高了抑郁筛查能力。虽然多模态识别正在成为抑郁症识别任务中的重点研究对象,但现有的多模态抑郁症识别模型仍存在许多问题,例如特征提取时语句间关联性较弱,不同模态进行特征融合较为随意和在中文数据集上模型的泛化能力缺乏验证等。

针对上述不足,本文提出了一种句子级时序卷积记忆网络的多模态抑郁症识别模型。该模型是以传统TCN模型为基础,在句子级特征提取中引入了用于捕捉序列中长期依赖信息的GRU门控单元[16]模块,这解决了TCN模型在特征提取时,语句间关联性较弱的问题,使每例数据的上下文语句构建起关联。为了能够直接关注序列中不同位置的相关性,以更好地捕捉长距离依赖关系又引入了Self-Attention,它可以将重要的位置与较不重要的位置进行区分,并在注意力计算过程中赋予不同位置不同的权重,使模型能够聚焦于关键的部分,从而增强了模型的表达能力。同时,在全局特征提取中不同模态特征值之间使用了Cross Attention,为不同模态特征值赋予不同的权重进行特征融合,并分别在DAIC-WOZ公开数据集和MMD2022私有中文抑郁症数据集上进行实验验证,结果表明,在多模态抑郁症识别任务上,STCMN模型各实验指标均得到提升,同时解决了多模态抑郁症识别模型在中文数据集上泛化能力不足的问题。

1 模型构建

1.1 TCN模型概述

在TCN模型中引入因果卷积、扩张卷积和残差块,在提取时序特征时既能控制模型的感受野大小又能将模型的参数数量控制在可接受范围内,使模型不会过于复杂。下面对每个结构进行详细描述。

1.1.1 因果卷积

因果卷积是TCN的关键架构。对于一维时间序列输入$X=(x_0,x_1,\cdots,x_t,\cdots,x_T)$,时刻$t$的输出$y_t$仅依赖于当前时间$x_t$和部分过去时间的输入(即$x_{t-1},x_{t-2},x_{t-3}$),而不依赖于任何未来时间的输入(即$x_{t+1},x_{t+2},x_{t+3},\cdots,x_T$)。这使网络的输出信息只会受过去输入信息的影响,从而避免了未来信息向过去信息的“泄露”。此外,因果卷积很容易受到感受野的限制,在卷积层较少时,模型的感受野较小,因此只能从较近的历史数据中提取信息来对当前时刻进行预测[17]。图1为因果卷积层堆栈的可视化。

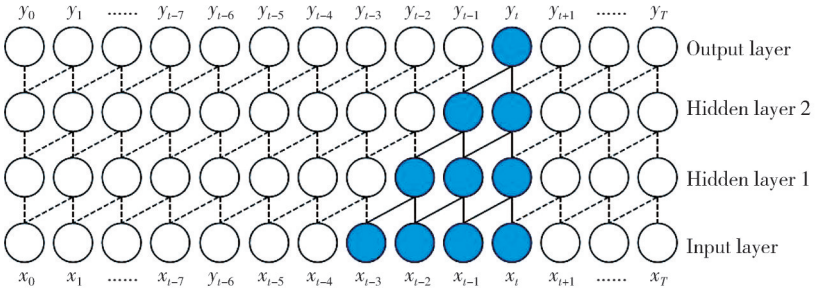

图1 因果卷积层堆栈的可视化

Fig. 1 Visualization of a stack of causal convolutional layers

1.1.2 扩张卷积

为解决因果卷积中的感受野受限问题,TCN模型在因果卷积的基础上引入了扩张卷积。对于一维时间序列,输入$X=(x_0,x_1,\cdots,x_t,\cdots,x_T)$和一个滤波器$f:\{0,1,2,\cdots,n-1\}$,序列元素$T$的扩张卷积运算定义见式(1)。

$$H(T)=(X*_d f)(T)=\sum_{i=0}^{n-1} f(i)\cdot x_{T-d\cdot i}, \quad (1)$$

式中:n为滤波器大小;d为膨胀因子;$T-d\cdot i$为过去的方向。通过增加滤波器大小n和膨胀因子d,TCN可以有效地扩大感受野,这使顶层的输出能够接收更广泛的输入信息。此外,通过在每层中并行处理相同的滤波器,也可以提高整个模型的计算效率。图2为扩张卷积层堆栈的可视化,此时,滤波器大小$n=2$,膨胀因子$d=[1,2,4]$。添加扩张卷积后,在时间t时刻的输出y_t可以接收的输入范围为:$x_{t-7},x_{t-6},\cdots,x_t$。

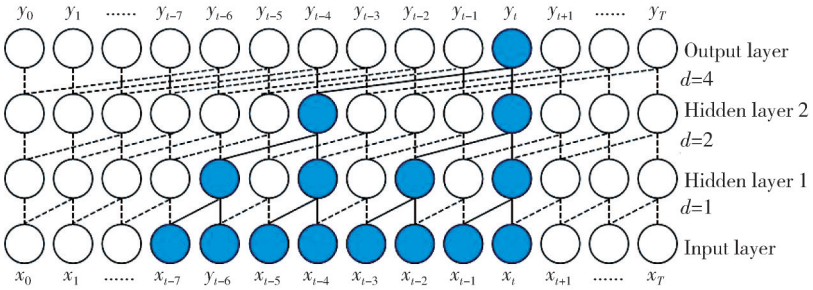

图 2　扩张卷积层堆栈的可视化

Fig. 2　Visualization of a dilated stack of convolutional layers

1.1.3　残差块

除了调整滤波器大小n和膨胀因子d外，还可以通过增加隐藏层的层数来扩大TCN模型的感受野。然而，非常深的网络会影响模型训练的稳定性，并出现梯度消失。为了解决这个问题，TCN模型中采用了残差块结构。残差块的细节如图 3 所示。

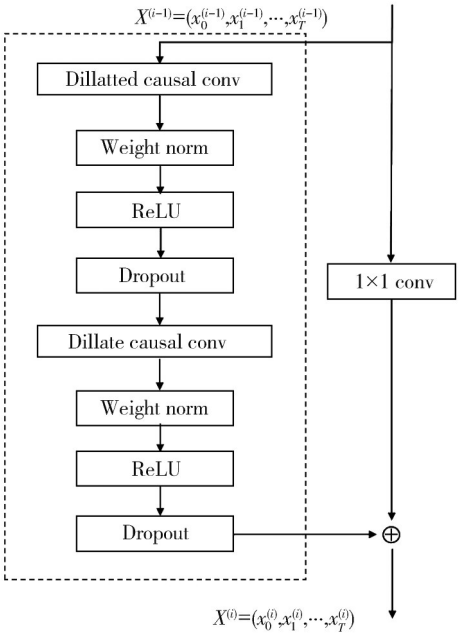

图 3　残差块内部结构

Fig. 3　Internal structure of the residual block

第i个残差块以上一个残差块的输出$X^{(i-1)}$作为输入，其中一个分支进行扩张卷积操作，另一分支使用1×1的卷积核对原始输入数据进行升维或降维操作，以便能够与卷积后的数据直接相加后得到输出$X^{(i)}$，计算公式为

$$X^{(i)}=\delta(F(X^{(i-1)})+X^{(i-1)}), \tag{2}$$

式中：δ为激活当前第i个残差块；F为残差块中的一系列操作，具体包括扩张卷积从输入中提取隐藏特征，权重范数通过限制范围来提高训练速度，激活层使用效果较好的ReLU函数作为激活函数，最后使用Dropout进行正则化来解决网络的过拟合问题。

1.2　STCMN总体模型的提出与构建

本研究通过将句子级多模态TCN模型与GRU、Self-Attention和Cross Attention相融合来提高对抑郁的识别能力，并提出STCMN（Sentence-level Temporal Convolutional Memory Networks）多模态深度学习模型，该模型利用时序的面部特征、患者音频特征和访谈转录文本特征作为模型的输入，根据句子级时间戳，对患者的每句话进行提取，再将每句话的不同模态特征数据分别传入对应模型单元进行句子级特征提取，然后使用GRU门控单元和Self-Attention对输出的特征进行上下文关联和加权组合，并捕捉序列中不同语句之间的长距离依赖关系，之后使用卷积神经网络将每句话的不同模态特征进行特征融合。为了降低模型复杂度和提升模型的时效性，提出了可变阈值M来表示语句数量的上限值，在获取多模态句子级特征后，将特征值传入全局特征提取单元，将最终得到的不同模态特征通过使用Cross Attention进行特征融合后传入全连接层以获得最后的抑郁识别标签。总体模型如图 4 所示。

特别的是，本文不仅提出了句子级多模态特征提取总体模型结构，还在模型中的句子级特征提取中引入特征提取单元TCN-GS（Temporal Convolutional Network-GRU-Self attention），并在模型进行全局特征提取后使用crossATT（Cross Attention）模块对不同模态特征进行融合。TCN-GS和crossATT组合不仅使模型在长序列的处理能力、上下文关联能力得到了提升，还从不同模态的信息中获取了更准确的特征表示，更进一步提高了模型的计算效率和抑郁症识别精度。

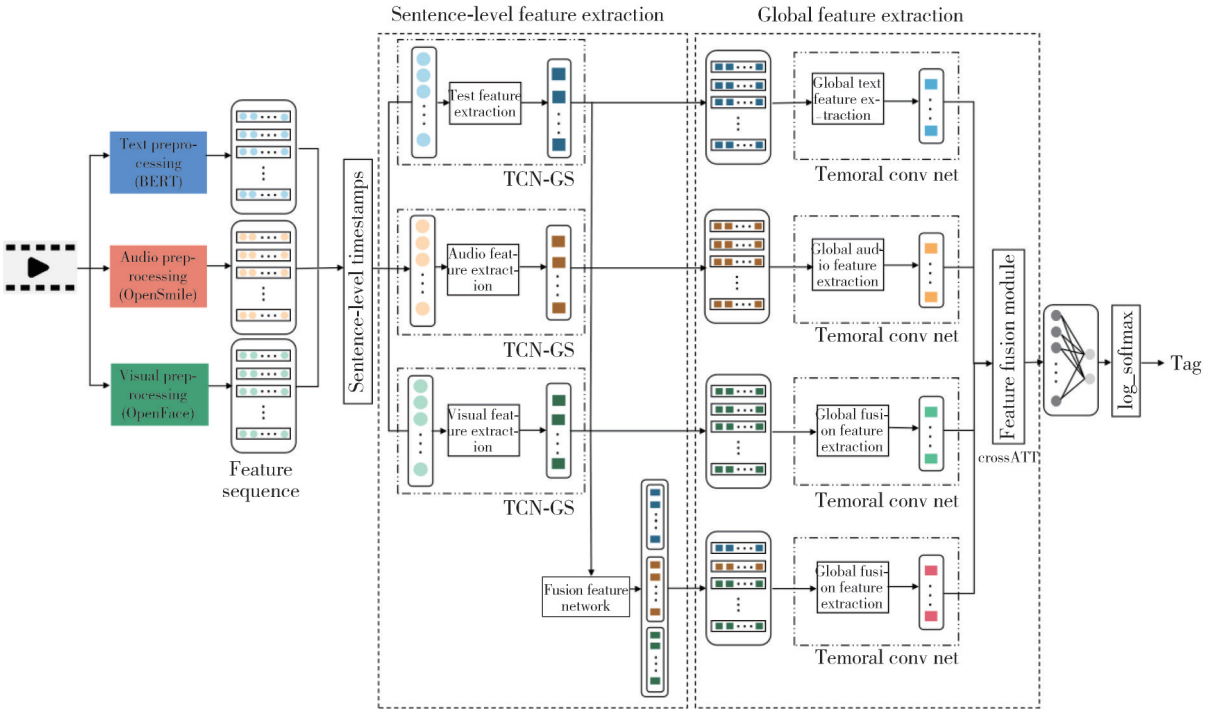

图4 本文提出的STCMN模型架

Fig. 4 The proposed STCMN model architecture

1.2.1 句子级特征提取模块

本文提出的TCN-GS模块是在原有TCN中残差块结构的基础上，又融合了GRU和Self-Attention模块，将患者语句特征之间产生关联，从而弥补了TCN模型缺乏上下文关联能力的局限性，GRU能通过循环机制和门控单元动态地更新和维护隐藏状态，对特征能够进行有效的整合和建模，二者相结合，可以更加全面地表示时间序列数据中的长期依赖关系，从而提高模型的预测能力，具体模块如图5所示。

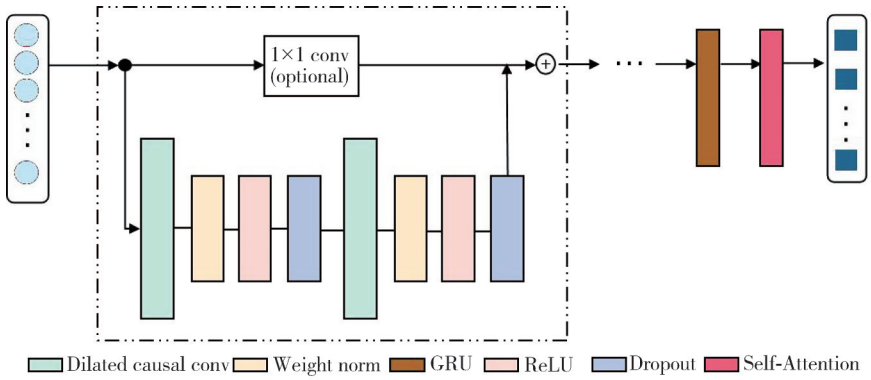

图5 TCN-GS句子级特征提取模块

Fig. 5 TCN-GS sentence-level feature extraction module

TCN-GS特征提取单元中加入GRU门控单元而不是使用效果更稳定的长短期记忆网络LSTM，主要原因是本文提出的句子级抑郁症识别总模型STCMN相比于传统的端对端TCN模型，结构更加复杂、模型参数更多，若要结合LSTM算法会使模型的时效性较差，要求的实验条件更加严格，因此，选用与LSTM算法类似的结构较为简单的GRU门控单元来满足该模型要求。如图6所示，X_t为本句特征序列，h_{t-1}为上一句的隐藏状态，h_t为要传递为下一句的隐藏状态。

GRU中主要有两个门结构：重置门r_t，这里要将上句的隐藏状态h_{t-1}与当前语句特征x_t按一定权重进行结合，见式(3)，其中$\boldsymbol{W}_t$并不是一个值而是一个权重矩阵，得到的值越小，需要遗忘的上一句特征就越多，有助于捕捉时间序列里短期的依赖关系。

$$r_t = sigmoid(\boldsymbol{W}_t \cdot [h_{t-1}, x_t])。 \tag{3}$$

更新门z_t，决定到底要将多少过去的信息传

递到未来，通过式（4）得到z_t，z_t越接近1，代表“记忆”下来的数据越多，而越接近0则代表“遗忘”的越多，这有助于捕捉时间序列中的长期依赖关系，$\boldsymbol{W}_z$代表权重矩阵。

$$z_t = sigmoid(\boldsymbol{W}_z \cdot [h_{t-1}, x_t])。\tag{4}$$

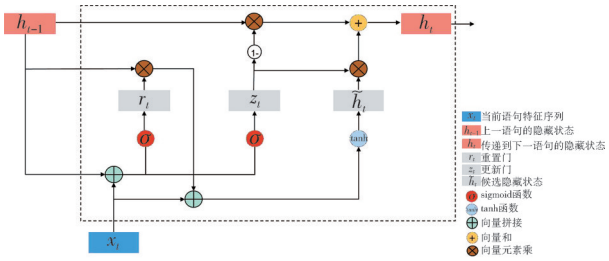

图 6　GRU门控单元

Fig. 6　GRU gating unit

除了主要的两个门结构，还要计算候选隐藏状态$\tilde{h}_t$，其中$\boldsymbol{W}$代表权重矩阵，更新记忆后传递给下一节点的隐藏状态h_t。

$$\tilde{h}_t = \tanh(\boldsymbol{W} \cdot [r_t \otimes h_{t-1}, x_t]),\tag{5}$$

$$h_t = (1 - z_t) \otimes h_{t-1} + z_t \otimes \tilde{h}_t,\tag{6}$$

式中：$\otimes$为向量元素乘；$(1 - z_t) \otimes h_{t-1}$为对原本隐藏状态的选择性“遗忘”；$z_t \otimes \tilde{h}_t$为对包含当前节点信息的选择性“记忆”。

1.2.2　全局多模态特征融合模块

CrossATT特征融合模块，通过交叉注意力机制（Cross Attention，CA）计算出F_{fuse}分别与F_{text}、F_{audio}、F_{face}之间的关系映射，将归一化的模态映射后再与F_{fuce}相乘，形成交叉注意力特征$F_{\text{text}}^{\text{CA}}$、$F_{\text{audio}}^{\text{CA}}$以及$F_{\text{face}}^{\text{CA}}$。最后，再将原始的$F_{\text{fuse}}$特征和交叉注意力特征拼接。在进行全局特征融合时，考虑到在多模态模型训练中容易被部分模态特征所主导而忽略其他特征，因此，STCMN模型是以融合特征F_{fuse}为主，将其与不同模态特征相结合来赋予不同的权重，保证text、audio和face作为fuse的辅助。如图7所示。

1.2.3　损失函数

损失函数是用来衡量模型预测结果与真实标签之间差异的函数，在STCMN模型中，使用的是在分类任务中常用的交叉熵损失函数（Cross Entropy Loss Function）。在二分类任务中，模型最后需要预测的结果只有两种情况，对于每个类别预测得到的概率为p和$1-p$，表达式为

$$L = -\frac{1}{N}\left[\sum_{i=1}^{N} y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right],\tag{7}$$

式中：N为样本数量；y_i表示样本i的真实标签值，正类为1，负类为0；p_i表示样本i预测为正类的概率。

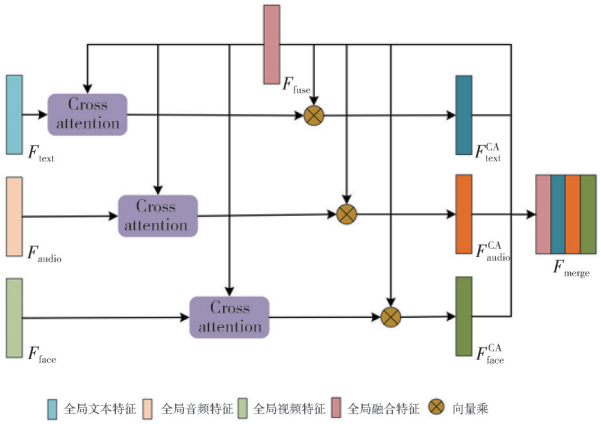

图 7　crossATT特征融合模块

Fig. 7　crossATT feature fusion module

2　多模态数据特征

本节将具体介绍本文提出的STCMN模型使用的多模态（音频、视频和文本）特征输入数据。其中，视频模态的具体输入特征见表1；音频模态的具体输入特征见表2；在使用Bert模型进行文本特征提取之前，使用表3中的特征对Bert模型进行微调，使其能够更好地区分抑郁症相关文本。

表 1　STCMN模型具体输入的面部特征

Tab. 1　Specific input facial features for the STCMN model

特征分类	面部特征	特征描述
3D面部特征点坐标	$x_0, x_1, x_2, x_3, \cdots, x_{67}, y_0, y_1, y_2, y_3, \cdots, y_{67}, z_0, z_1, z_2, z_3, \cdots, z_{67}$	坐标点在世界坐标空间中以毫米为单位，摄像头位于(0, 0, 0)
面部动作单元(AU)	AU01_r, AU02_r, AU04_r, AU05_r, AU06_r, AU09_r, AU10_r, AU12_r, AU14_r, AU15_r, AU17_r, AU20_r, AU25_r, AU26_r, AU04_c, AU12_c, AU15_c, AU23_c, AU28_c, AU45_c	后缀是“_r”表示的值是对应序号的动作单元的回归输出；后缀是“_c”反映对应序号的动作单元的二进制标签，存在为1，不存在为0
眼睛聚焦方向特征	$x_0, y_0, z_0, x_1, y_1, z_1, x_{h0}, y_{h0}, z_{h0}, x_{h1}, y_{h1}, x^{h1}$	前两个向量表示在世界坐标空间中两只眼睛的聚焦方向，后两个向量表示头部的空间坐标
头部姿势	$T_x, T_y, T_z, R_x, R_y, R_z$	T_x, T_y, T_z表示位置坐标；R_x, R_y, R_z表示头部旋转坐标，位置在世界坐标空间中以毫米为单位

表2 STCMN模型具体输入的音频特征

Tab. 2 Specific input audio features for the STCMN model

特征分类	音频特征	特征描述
F0 (Fundamental Frequency)特征	平均基频(Mean F0)	捕捉声音信号的音高或音调
	基频标准差(Standard deviation of F0)	表示情感表达的不稳定性
	最小基频(Minimum F0)、最大基频(Maximum F0)	识别声音信号中的音调范围
HNR (Harmonic-to-Noise Ratio)特征	平均 HNR(Mean HNR)	检测声音清晰度的变化
	HNR 标准差(Standard deviation of HNR)	衡量声音信号的稳定性
光谱斜度 (Spectral Slope)特征	平均光谱斜度(Mean spectral slope)	描述声音信号的频谱倾斜度
	光谱斜度标准差 (Standard deviation of spectral slope)	检测频谱特性变化的不规则性
MFCCs (Mel-Frequency Cepstral Coefficients)特征	梅尔频率倒谱系数的均值(Mean of MFCCs)	提供了声音信号的频谱信息
声学环境特征(Acoustic Environment Features)	声噪比的均值 (Mean of Signal-to-Noise Ratio) 声噪比标准差 (Standard deviation of Signal-to-Noise Ratio)	患者可能受到的声音环境

表3 微调Bert模型使用的文本特征

Tab. 3 Fine-tuning the text features used by the Bert model

特征分类	特征描述	举例
情感表达	负面情绪	“悲伤”“沮丧”“绝望”“孤独”“压抑”等
	情感强度	“痛苦”“无助”“不值得”等
自我评价	自我贬低	“我很失败”“我一无是处”“自己是个彻底的失败者”“我一点价值都没有”等
	自责	“我害了我的家人，没有我他们会更好。”“每次出错都让我感到内疚，我觉得我对别人造成了困扰。”等
生理症状	疲劳和精力不足	“每天睡很久但还是感觉困”“干什么都觉得很累，总想睡觉”等
	睡眠问题	“晚上躺床上很长时间都睡不着”“有时凌晨醒来就很难再睡着”“睡眠浅，很容易醒还容易做噩梦”等
	食欲变化	“没胃口，不想吃”“自己以前喜欢的东西也不想吃”等
	消化问题	“胃痛”“胃胀”“便秘”“腹泻”等
	慢性疼痛	“慢性头痛”“肌肉疼痛”“背部疼痛”等
社交问题	孤立感	“我不想见人”“不愿意和别人交往”等
	社交回避	“不想和朋友、家人互动”“害怕与别人交流”“不愿意接打电话”“不想回消息”等
自杀倾向	自杀念头	“我感到生活毫无意义，我宁愿死了。”等
	自杀计划	“自己写过遗书”“吃过安眠药但没死成”等

3 实验与讨论

3.1 数据集

本文使用的数据集有：公开数据集DAIC-WOZ和私有数据集MMD2022，私有数据来自项目合作医院的多模态抑郁症数据库。其中，DAIC-WOZ包含了非抑郁人群和抑郁患者的音频、视觉特征和转录文本，并为每例数据进行了PHQ-8评分。私有数据集MMD2022为正在构建的中文抑郁症数据库，是以半结构化的临床访谈形式创建的。在这些访谈中，患者与远程控制的数字化身进行交谈，临床医生通过数字化身提出一系列用于评估患者抑郁水平的问题，这通常会包括一些敏感性问题(例如：“是否会感到自己很糟糕，觉得自己很失败，让自己或家人失望”)和一些模拟对话反馈(例如：“嗯”)，使访谈场景更加真实。该数据集现已收集处理了98例临床访谈数据，原视频数据约20 h，且每例数据均有对应的汉密尔顿抑郁量表(HAMD-17)、蒙哥马利抑郁评定量表(MARDS)和汉密尔顿焦虑量表(HAMA)评分结果。需要强调的是，MMD2022数据集中不包含参与者的私人信息，原视频文件仅用于进行多模态数据特征提取，数据集中仅存放提取出的特征数据文件，参与者的音频和转录文本中提到的个人姓名、特定日期和地点均已删除。面部扫描的分辨率也较低，不能够用来识别个体，但包含了足够用于识别的面部特征。

3.2 预处理

本文所用到的DAIC-WOZ数据集中包括了视觉特征、音频、转录文本和量表评分文件。构建MMD2022中文抑郁数据库的原数据为参与者的访谈视频。但是STCMN模型并不是直接将视频文件作为输入，而是先提取出视频的音频和转录文本文件，再将不同类型的数据经过不同的预处理操作后得到多维度特征向量或特征点坐标，并将其作为模型的输入。

3.2.1 时序面部特征提取

图像数据的预处理步骤一般包括给定的视觉输入、照明标准化、图像序列配准和对齐以及人脸检测。OpFace工具被广泛用作图像预处理工具，它是基于Viola等[18]提出的人脸检测算法的面部表情分析应用程序。本文利用OpenFace工具将每个参与者

的视频文件根据句子级时间戳在原视频中截取对应语句的视频片段，并以0.033 s为时间步长提取面部特征数据，如特征点坐标、头部姿势、眼睛凝视角和面部动作单元等，如图 8 所示。

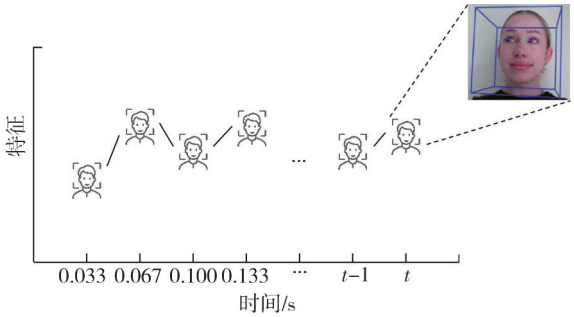

图 8　基于OpenFace工具的面部特征提取

Fig. 8　Extraction of facial features based on the OpenFace tool

3.2.2　音频特征提取

音频数据为每位患者完整的访谈音频，在提取音频特征之前要先按照该患者转录文本中的句子级时间戳将每句话音频切分提取，再通过音频特征提取工具得到特征数据。本研究中选用OpenSmile音频特征提取工具，使用其中的eGeMAPS音频特征集。对于抑郁症识别，抑郁症患者的语音通常会展现出与情感相关的特征，如语调变化、声音强度变化等，该特征集可以更好地捕捉这些与情感相关的音频特征，且相对于传统的声学特征集，它包含了更多维度的特征，如基音频率、频谱特性、梅尔频率倒谱系数等，更有助于区分抑郁症患者和非抑郁症人群之间的声音差异，如图 9 所示。

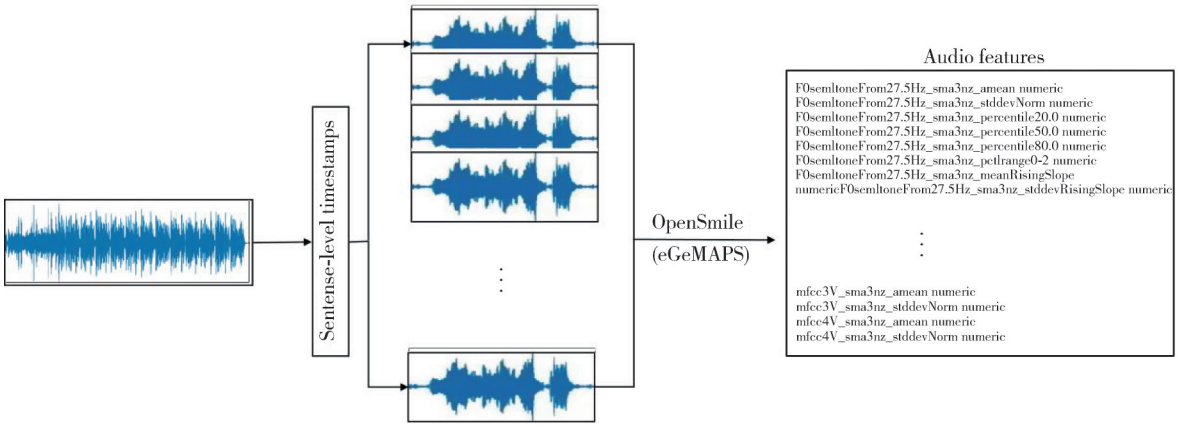

图 9　基于OpenSmile工具的音频特征提取

Fig. 9　Extraction of audio features based on the OpenSmile tool

3.2.3　文本特征提取

本文中的文本数据是由患者访谈音频转录得到的访谈文本。为了便于之后的中文抑郁症识别研究，在使用DAIC-WOZ数据集时需要将已有的转录文本翻译为中文。由于中英文语言差异和翻译工具的原因，翻译后的语句可能会出现没有意义的特殊字符或内容，如：“￥”或“＜清嗓子＞”等，需要将这些内容处理后才能使用。重要的是，并不是将所有文本内容作为模型输入，而是根据身份标签，只提取患者语句。最后，使用Bert模型[19]对患者语句进行文本特征提取，得到文本特征文件，如图 10 所示。

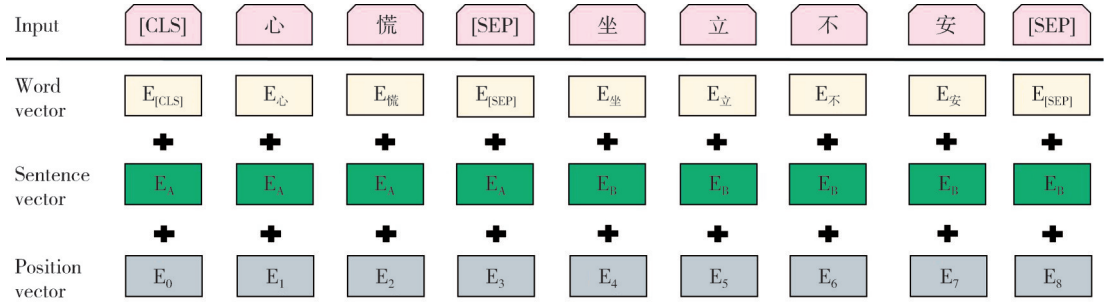

图 10　基于BERT模型的文本特征提取

Fig. 10　Extraction of text features based on the BERT model

3.3　实验环境和参数设置

本文实验基于Linux操作系统，训练采用的硬件设备主要为3块GeForce RTX 3090 GPU，软件环境包括python3.6.6和pytorch1.7.0框架。采用时域卷积网络TCN作为基础网络，在对其优化后

得到句子级时序卷积记忆网络STCMN，该模型的主要参数设置如下：初始学习率为1×10^{-3}，优化器为Adam，句子级特征提取中隐藏层数为5，隐藏层通道数为12，卷积核大小为5，dropout为0.5。

3.3.1 实验数据

本文根据两个数据集在不同模型中分别进行实验。DAIC-WOZ数据集中共有189例数据。但可用数据共有142例，其中以114例作为训练数据，28例作为测试数据。MMD2022数据集共有98例数据，可用数据为92例，其中以74例作为训练数据，18例作为测试数据。本文以抑郁症样本作为正样本，正常样本作为负样本。在作为模型的输入数据之前，由于训练集中的正负样本数量并不相同，因此为了保证模型选取训练数据时的正负样本概率相同，要对训练数据中的正负样本赋予不同的权重weight。数据平衡操作依据式(8)和式(9)进行训练数据集中正负样本占比R_{P}的计算。

$$R_{\mathrm{P}}=\left[\frac{N_0}{S_t},\frac{N_1}{S_t}\right], \tag{8}$$

$$V_{\mathrm{W}}=\frac{1}{R_{\mathrm{P}}}。 \tag{9}$$

式中：S_t表示训练集样本总数；N_0表示训练集中负样本的个数；N_1表示训练集中正样本的个数。

3.3.2 实验评价指标

本文使用精确率(Precision)R_{pre}、召回率(Recall)R_{rec}、特异性(Specificity)R_{spe}和$F1$-score值作为实验的评价指标来评估不同分类模型的性能。其中，精确率也叫查准率，表示在预测为正样本的所有样本中，真实的正样本所占的比例；召回率也叫查全率，表示所有真实的正样本中，被预测为正的样本所占的比例。在理想状态下，精确率和召回率越高，代表模型的性能越好，但是事实上两者是矛盾的，而$F1$值作为精确率和召回率的调和平均值，用来评估模型在精确率和召回率上的综合性能[20]。实验中特异性也是一个很重要的指标，特异性越高，误诊的比例越低。各个指标的计算公式分别为

$$R_{\mathrm{rec}}=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}}, \tag{10}$$

$$R_{\mathrm{pre}}=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}}, \tag{11}$$

$$F1=\frac{2N_{\mathrm{TP}}}{2N_{\mathrm{TP}}+N_{\mathrm{FP}}+N_{\mathrm{FN}}}, \tag{12}$$

$$R_{\mathrm{spe}}=\frac{N_{\mathrm{TN}}}{N_{\mathrm{TN}}+N_{\mathrm{FP}}}, \tag{13}$$

$$R_{\mathrm{acc}}=\frac{N_{\mathrm{TP}}+N_{\mathrm{TN}}}{N_{\mathrm{TP}}+N_{\mathrm{TN}}+N_{\mathrm{FP}}+N_{\mathrm{FN}}}, \tag{14}$$

式中：N_{TP}为被分类正确的正样本数量；N_{FP}为被分类成正样本的负样本数量；N_{TN}为被分类正确的负样本数量；N_{FN}为被分类成负样本的正样本数量。

3.3.3 隐藏层数对模型性能的影响

本文实验以DAIC-WOZ公开数据集中的114例数据作为训练数据，28例数据作为测试数据，且保证每组实验的训练数据目录和测试数据目录相同，使用了Adam优化器对每组实验进行最多100个epoch的训练，设置了初始学习率为1×10^{-3}，每5轮训练后学习率将变为原来的1/10，如果连续10轮测试集上的损失没有减少将停止训练。通过每组实验仅改变句子级特征提取中的隐藏层数的方法，来验证不同层数对STCMN模型效果的影响。实验共做3次且每次的训练数据目录和测试数据目录均不相同，最终取3次实验数据的均值作为实验结果，具体实验数据见表4。

表4 不同隐藏层层数下STCMN模型的$F1$值、特异性、精确率、召回率和准确率指标

Tab.4 $F1$ score, specificity, precision, recall and accuracy metrics of STCMN models with different number of hidden layers

实验序号	层数	$F1$ Score	Specificity	Precision	Recall	Accuracy
1	3	0.52	0.60	0.57	0.48	0.55
2	4	0.78	0.68	0.71	0.86	0.74
3	**5**	**0.92**	**0.94**	**0.94**	**0.90**	**0.91**
4	6	0.90	0.91	0.91	0.89	0.88
5	7	0.90	0.91	0.91	0.89	0.89
6	8	0.89	0.90	0.90	0.88	0.88

由表4可得，句子级特征提取中隐藏层数为5时STCMN模型的性能最好，模型的$F1$值、特异性、精确率、召回率和准确率均优于其他层数时的模型指标。

3.3.4 对比实验结果与分析

实验A使用DAIC-WOZ公开数据集对不同的模型在多模态(文本、音频和视觉)特征数据下的效果进行对比评估，实验结果见表5。

表 5　实验 A 中不同模型下多模态数据的 $F1$ 值、特异性、精确率、召回率和准确率指标

Tab. 5　$F1$ score, specificity, precision, recall and accuracy indicators of multimodal data in different models in experiment A

实验序号	模型	$F1$ Score	Specificity	Precision	Recall	Accuracy
1	CNN	0.67	0.70	0.65	0.71	0.69
2	GRU	0.65	0.68	0.69	0.63	0.66
3	LSTM	0.77	0.79	0.75	0.79	0.78
4	TCN(all)	0.76	0.78	0.71	0.83	0.77
5	TCN(sentence)	0.79	0.81	0.74	0.85	0.80
6	CNN-Attention	0.73	0.76	0.68	0.78	0.75
7	GRU-Attention	0.82	0.85	0.78	0.86	0.84
8	BiAttention-GRU	0.90	0.91	0.91	0.89	0.89
9	**STCMN**	**0.92**	**0.94**	**0.94**	**0.90**	**0.91**

根据上述实验结果，实验 1 与实验 2、实验 3 相比，CNN 要优于记忆网络 GRU，但不如更复杂的长短期记忆网络 LSTM，证明在多模态抑郁症识别研究中，使用记忆网络算法提高识别效果的可能性。实验 3 与实验 4、实验 5 相比，使用了因果卷积和残差连接的 TCN 模型，在整体特征作为输入模式下除了 Recall，其余指标均略差于 LSTM 模型，但句子级 TCN 模型除了 Precision 与 LSTM 模型略差 0.01，其余指标均高于 LSTM 模型，证明了本研究以句子级 TCN 模型为基础进行改进的可行性。实验 1 和实验 6 都使用了 CNN 算法，比较模型的不同指标，可以发现融入 Attention 后模型的识别精度会有所提高，而实验 2 和实验 7 也证明了该结论。实验 7 和实验 8 相比，证明基于 GRU 算法使用双注意力模型效果会更优。最后，本文提出的基于 TCN 算法并融合了 GRU、Self-Attention 和 Cross attention 的句子级多模态抑郁症识别模型 STCMN 在各项指标中均优于其他模型，$F1$ 值为 0.92，特异性为 0.94，精确率为 0.94，召回率为 0.90，准确率为 0.91。对比识别效果相对较好的 BiAttention-GRU 模型，本文模型在 $F1$ 值、特异性、精确率、召回率和准确率上分别提升了 0.02，0.03，0.03，0.01 和 0.02。实验结果表明，基于句子级 TCN 模型融入 GRU 算法和注意力机制后，对于多模态抑郁症的识别非常有效，能够降低误诊率，这对抑郁症的辅助诊断有很大的帮助。

实验 B 使用 MMD2022 私有数据集对不同的模型在多模态（文本、音频和视觉）特征数据下的效果进行对比评估，本次实验只将两种 TCN 模型、BiAttention-GRU 模型与本文提出的 STCMN 模型进行比较，实验结果见表 6。

表 6　实验 B 中多模态数据在不同模型的 $F1$ 值、特异性、精确率、召回率和准确度指标

Tab. 6　$F1$ score, specificity, precision, recall and accuracy indicators of multimodal data in different models in experiment B

实验序号	Model	$F1$ Score	Specificity	Precision	Recall	Accuracy
1	TCN(all)	0.89	0.86	0.89	0.90	0.85
2	TCN(sentence)	0.91	0.90	0.91	0.92	0.89
3	BiAttention-GRU	0.94	0.94	0.94	0.94	0.93
4	**STCMN**	**0.97**	**0.98**	**0.98**	**0.96**	**0.97**

由表 6 可知，本文提出的 STCMN 模型，在私有数据集 MMD2022 上相比其他模型仍有更好的效果，证明了 STCMN 模型在中文抑郁症识别任务上具有较好的性能和泛化能力，能够适应不同抑郁症数据集的特征，同时也证明了本研究中数据预处理阶段使用的视觉特征提取方法的可行性和特征参数选择的有效性，所以，STCMN 模型在解决多模态数据识别抑郁症问题上是一种更为有效的选择。

3.3.5　消融实验结果与分析

为了更好地验证本文所提模型的优越性和所加组件的合理性，在 DAIC-WOZ 公开数据集上进行了消融实验，并且在保证模型其他部分的超参数相同的前提下，对加入不同组件后的模型性能进行评估，实验结果见表 7。

在上述实验中，将端到端的整体 TCN 模型作为基线模型，模型性能指标见表 7 实验 1。为验证句子级结构对模型效果的影响，将基线模型改为端到端的句子级 TCN 模型，通过实验 2 证明了句子级结构对抑郁症识别任务有着较好的影响。实验 3 在句子级 TCN 模型的基础上融入了 GRU 门控单元，这使得模型的各项指标分别提升了 0.03，0.03，0.05，0.01 和 0.03，证明了 GRU 组件对模型性能提升的有效性。通过实验 3 和实验 4 指标数据对比可以看出，Self-Attention 的加入使该模型的性能得到了进一步提升。最后，在融合了 Cross Attention 的实验 5

数据中，得出此时模型的最终性能与实验4相比又分别提升了0.03，0.03，0.04，0.01和0.03，证明了模型融合Cross Attention的有效性，且与基线模型整体TCN模型相比，模型得到了较为明显的优化，模型性能分别提升了0.16，0.16，0.23，0.07和0.14。通过本实验可得，本文提出的STCMN模型中所融入的组件在抑郁症识别任务中均发挥了良好的效果。

表7 消融实验中添加不同组件后模型的$F1$值、特异性、精确率、召回率和准确率指标

Tab.7 $F1$ score, specificity, precision, recall and accuracy metrics of the model after adding different components in the ablation experiment

实验序号	模型	$F1$ Score	Specificity	Precision	Recall	Accuracy
1	TCN(all)	0.76	0.78	0.71	0.83	0.77
2	TCN(sentence)	0.79	0.81	0.74	0.85	0.80
3	TCN(sentence)+GRU	0.82	0.84	0.79	0.86	0.83
4	TCN(sentence)+GRU+Self-Attention	0.89	0.91	0.90	0.89	0.88
5	TCN(sentence)+GRU+Self-Attention+Cross Attention	**0.92**	**0.94**	**0.94**	**0.90**	**0.91**

4 总结与展望

本文在抑郁症识别领域TCN模型的基础上提出了一种句子级时序卷积记忆网络的多模态抑郁症识别模型STCMN。该模型将多模态特征数据作为输入，通过模型识别患者是否患有抑郁症。通过实验证明，本文提出的STCMN模型在实验评估模型的所有指标，无论是在公开的DAIC-WOZ数据集还是在正在构建的MMD2022私有数据集上均有提升。为了保证患者数据的私密性，公开的数据集中并不包含参与者的原视频文件，这在很大程度上限制了使用多模态数据对抑郁症识别的研究，尤其是在视觉模态上的研究。现公开的数据集多为英文数据集，因为中英文语法和参与者的生活环境等因素的不同，这也将导致在提取文本和音频特征时会存在一定的差异，最后的模型结果可能不会达到最优。此外，由于数据集中参与者数量较少和抑郁参与者的不平等分布，对模型效果会有影响并增加了额外的建模挑战。因此，我们正在构建一个参与者较多并包含有更多的视觉特征文件和多个由专业精神科医生评分的量表数据的中文抑郁数据库MMD2022，以进一步加强中文抑郁症识别的研究。为了保护参与者的私人健康信息，MMD2022数据集并不包含有包括原视频文件在内的所有有关参与者隐私的信息。目前，我们在多模态抑郁症识别领域还处在实验探索阶段，本文所提出的STCMN模型是二分类模型，并不能识别出抑郁等级，因此，后续将紧跟技术前沿不断优化完善MMD2022数据集中的特征信息和探索新的更加有效的多分类抑郁症识别模型。

参考文献：

[1] SANTOMAURO D F, HERRERA A M M, SHADID J, et al. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic[J]. Lancet, 2021, 398(10312): 1700-1712.

[2] MCGINNIS E W, ANDERAU S P, HRUSCHAK J, et al. Giving voice to vulnerable children: machine learning analysis of speech detects anxiety and depression in early childhood[J]. IEEE Journal of Biomedical and Health Informatics, 2019, 23(6): 2294-2301.

[3] DI MATTEO D, FOTINOS K, LOKUGE S, et al. The relationship between smartphone-recorded environmental audio and symptomatology of anxiety and depression: exploratory study[J]. JMIR Formative Research, 2020, 4(8): e18751.

[4] FLORES R, TLACHAC M L, TOTO E, et al. Transfer learning for depression screening from follow-up clinical interview questions[M]. Singapore: Springer Nature Singapore, 2022: 53-78.

[5] TLACHAC M L, RUNDENSTEINER E. Screening for depression with retrospectively harvested private versus public text[J]. IEEE Journal of Biomedical and Health Informatics, 2020, 24(11): 3326-3332.

[6] SENN S, TLACHAC M L, FLORES R, et al. Ensembles of bert for depression classification[C]// 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2022: 4691-4694.

[7] AL JAZAERY M, GUO G. Video-based depression level analysis by encoding deep spatiotemporal features[J]. IEEE Transactions on Affective Computing, 2021, 12(1): 262-268.

[8] WANG Q, YANG H, YU Y. Facial expression

video analysis for depression detection in Chinese patients [J]. Journal of Visual Communication and Image Representation, 2018, 57: 228-233.

[9] ASGARI M, SHAFRAN I, SHEEBER L B. Inferring clinical depression from speech and spoken utterances [C]//2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2014: 1-5.

[10] RODRIGUES MAKIUCHI M, WARNITA T, UTO K, et al. Multimodal fusion of bert-cnn and gated cnn representations for depression detection [C]//9th International Conference on Audio/Visual Emotion Challenge and Workshop, 2019: 55-63.

[11] TOTO E, TLACHAC M L, Rundensteiner E A. Audibert: A deep transfer learning multimodal classification framework for depression screening [C]//30th ACM International Conference on Information & Knowledge Management, 2021: 4145-4154.

[12] HAQUE A, GUO M, MINER A S, et al. Measuring depression symptom severity from spoken language and 3D facial expressions [DB/OL]. (2018-11-21) [2023-09-27]. https://arxiv.org/abs/1811.08592.

[13] RAY A, KUMAR S, REDDY R, et al. Multi-level attention network using text, audio and video for depression prediction [C]//9th International Conference on Audio/Visual Emotion Challenge and Workshop, 2019: 81-88.

[14] CAO Y, HAO Y, LI B, et al. Depression prediction based on BiAttention-GRU [J]. Journal of Ambient Intelligence and Humanized Computing, 2022, 13 (11): 5269-5277.

[15] FLORES R, TLACHAC M L, TOTO E, et al. AudiFace: Multimodal deep learning for depression screening [C]//Machine Learning for Healthcare Conference. PMLR, 2022: 609-630.

[16] CHUNG J, GULCEHRE C, CHO K, et al. Gated feedback recurrent neural networks [C]//International Conference on Machine Learning, PMLR, 2015: 2067-2075.

[17] WANG Y Y, CHEN J, CHEN X Q, et al. Short-term load forecasting for industrial customers based on TCN-LightGBM [J]. IEEE Transactions on Power Systems, 2021, 36(3): 1984-1997.

[18] VIOLA P, JONES M J. Robust real-time face detection [J]. International Journal of Computer Vision, 2004, 57(2): 137-154.

[19] MARTÍNEZ-CASTAÑO R, HTAIT A, AZZOPARDI L, et al. Early risk detection of self-harm and depression severity using BERT-based transformers [C]//Proceedings of the Working Notes of CLEF, 2020.

[20] 刘豪, 卓广平, 乔俊福, 等. 基于领域情感词典与字词特征融合的中文抑郁症文本分类方法[J]. 中北大学学报(自然科学版), 2022, 43(6): 522-529.
LIU Hao, ZHUO Guangping, QIAO Junfu, et al. Chinese depression text classification based on domain emotion dictionary and word feature fusion [J]. Journal of North University of China (Natural Science Edition), 2022, 43(6): 522-529. (in Chinese)