文章编号：1003－1251(2025)03－0024－08

基于时空差异网络的抑郁检测方法研究

吴东升，徐鹏飞，林玉婷

（沈阳理工大学 自动化与电气工程学院，沈阳 110159）

摘　要：在面部表情识别领域，针对不同抑郁水平的面部表情差异小以及模型计算复杂度高导致过度拟合的问题，提出一种由对角注意力机制(DMA)和时空差异模块(SDN)相结合的深度学习架构(DSN)进行抑郁检测，并根据面部表情的变化进行抑郁症等级评估。引入DMA使模型具备精确识别图像之间变化的能力；使用SDN模块捕获平滑与突然的面部表情变化，探索多尺度时间信息。实验结果表明，本文提出的DSN架构在保证模型性能的同时相对于传统架构降低了42%的运算复杂度，提高了模型整体容错水平，具有较好的泛化能力。

关　键　词：面部表情识别；抑郁检测；时空差异模块；深度学习

中图分类号：TP391.4　　**文献标志码：**A　**DOI：**10.3969/j.issn.1003－1251.2025.03.004

Research on Depression Detection Based on Spatiotemporal Difference Networks

WU Dongsheng，XU Pengfei，LIN Yuting

（Shenyang Ligong University，Shenyang 110159，China）

Abstract：In the field of facial expression recognition，to address the challenges of minimal facial expression differences at different depression levels and high computational complexity leading to overfitting，a deep learning architecture(DSN) is proposed for depression detection，which combins a diagonal attention mechanism(DMA) and a spatiotemporal differences network(SDN) to assess depression levels based on facial expression variations. The introduction of DMA enables the model to accurately identify changes between images. Meanwhile，the SDN module captures smooth and abrupt facial expression changes，exploring multiscale temporal information. Experimental results indicate that the proposed DSN architecture not only ensures model performance but also reduces computational complexity by 42% compared to traditional architectures. This enhancement improves the overall fault tolerance level of the model，demonstrating its superior generalization capabilities.

Key words：facial expression recognition；depression detection；spatiotemporal differences network；deep learning

抑郁症作为一种情绪障碍，对个人的精神健康构成严重威胁。抑郁症患者常常表现出积极情绪减退、自我意识降低和认知能力减弱等特征，易出现自残行为和自杀倾向[1-2]。传统的诊断方法通常根据患者的主观报告和医生的临床观察，若患者不愿或无法准确描述自己的状况，会给医务

收稿日期：2024－02－20

基金项目：辽宁省教育厅高等学校重点攻关项目(JYTZD2023006)

作者简介：吴东升(1973—)，男，教授，博士。

人员带来诊断困难,导致患者因得不到及时治疗而病情恶化。

作为一种替代方法,自动抑郁检测(ADD)旨在通过客观、可靠和自动化的方式,如生物标志物研究技术、移动健康技术、虚拟现实技术和人工智能技术等,收集受试者的脑电图(EEG)[3]、功能性近红外光谱(fNIRS)[4]、脑成像、眼球运动、面部表情[5-6]和行为线索等生理或行为信号作为检测依据。上述信号中,面部表情具有易于获取和非接触测量的优点,同时与抑郁症保持着密切的相关性[7]。因此,预测抑郁程度广泛采用的方式是分析受试者视频中的面部信息。

随着深度学习(DL)技术的发展,多种基于视频面部分析的网络架构被用来进行抑郁检测。Jan 等[8]采用预先训练好的 VGG 网络探索空间信息,并通过特征动态历史直方图(FDHH)获取时间信息,从而预测样本的抑郁程度。Zhou 等[9]提出的 DepressNet 是一种四流网络,可将人脸大致分为三个部分(即顶部、中部和底部),与原始图像一起作为输入数据,从而预测每帧图像的贝克抑郁量表(BDI-II)分数,由计算 BDI-II 分数的平均值得出相应的视频预测结果。Zhou 等[10]通过在 3D-CNN 框架中加入联合标签分布和度量学习,实现了时空信息表示的增强。He 等[11]使用 3D CNN 检测不同抑郁程度个体的面部变化。Niu 等[12]引入了一种新型的时空注意力(STA)网络,利用注意力机制整合空间和时间信息预测样本的 BDI-II 分数。同时,Niu 等[13]提出了一种利用图形卷积嵌入(GCE)和多尺度矢量化(MSV)技术的抑郁症预测模型。上述方法用来深入研究通道关系,并获得更精细的抑郁症表现形式。

尽管已提出几种基于面部视频数据进行抑郁检测的 DL 模型[14-15],仍难以维持高性能,其原因主要有两个。首先,DL 模型未能有效编码面部表情的时空特征,导致模型对具有不同标签的视频序列中实际存在的表情动态差异区分能力不足。因此,使用更强调区分性的时空信息模型[16]有助于提高表征的能力。其次,可用于设计预测架构的标注训练数据量有限,而训练数据集规模相对较小时会存在过拟合的风险。如果使用三维架构捕捉时空特征[17],通常需要以高计算复杂度为代价获得高性能。虽然可以使用不同形式的时空卷积以减少三维模型的计算成本,但这种方法主要关注固定的时空信息[18],从而降低了用于抑郁检测的具有区分性特征的表征潜力。

为了解决上述方法可能存在的问题,本文提出一种基于时空差异网络(SDN)的有效架构,用于探索不同时间尺度下的面部表情变化,可在有限的训练数据下学习获得丰富的面部表情变化,并通过检测面部表情特征来预测患者的 BDI-II 分数。该模型包含对角注意力机制(DMA)和 SDN:DMA 可以增强模型对面部特征细节的学习能力,进而提高模型发现和识别图像间所有变化的能力;SDN 由最大化区块和差异化区块组成,最大化区块用于捕捉面部结构平滑过渡,而差异化区块则编码面部突然的时空变化,利用不同区块的优势,突出相关要素,高效地探索面部各种时空变化信息。这些区块不依赖于三维滤波器,生成的特征组合方式为抑郁检测提供了稳健的特征表示。本文提出的抑郁检测方法简称为 DSN 方法。

1　对角注意力机制

注意力机制模拟人类在处理信息时的关注机制[19],使模型能够有选择性地集中关注输入数据中的特定部分,从而更有效地处理重要信息。在实际应用中,选择和设计合适的注意力机制需要考虑具体问题和数据的特征,以适应不同领域和任务的需求[20],提高模型在处理复杂和动态数据时的适应性,并发挥其捕捉长距离依赖关系、提供上下文信息以及处理不同位置信息方面具有的显著优势[21]。然而,自注意力机制计算复杂度较高且难以处理大规模数据,是需要应对的挑战。

本文综合自注意力机制的优势与挑战,提出 DMA 机制,在处理序列数据时只关注相互关联元素对角线上的序列,较大程度地减少了计算量,提高了模型性能。DMA 的详细信息如图 1 所示,其处理过程说明如下。

将输入帧 $I_t \in \mathrm{R}^{A\times M\times O}$ 定义成多个不重叠的补丁 p_t,如式(1)、式(2)所示。

$$\boldsymbol{p}_t = \{\boldsymbol{p}_t^i\}_{i=0}^{D_p} \tag{1}$$

$$|\boldsymbol{p}_t| = AM/(ps)^2 \tag{2}$$

式中:A 为通道高度;M 为通道宽度;O 为通道数量;ps 为每个补丁 p_t^i 的边长大小;D_p 为补丁块的数量。

首先对输入图像进行块交换操作,即在两帧图像之间随机选取一个相同的区域块;然后交换相对应的区域块 $\boldsymbol{p}_t^i$ 和 $\boldsymbol{p}_{t+\delta}^i$,再用这个对应区域块创建出交换图像 $\boldsymbol{I}_{t/s}$;最后使模型发现图像的变化并根据 $\boldsymbol{I}_{t/s}$ 重建 $\boldsymbol{I}_t$。模型还需学习如何从 $\boldsymbol{P}_{t/s}$ 重建

$\boldsymbol{p}_{\mathrm{t}}$,因为 $\boldsymbol{p}^{i}_{t\prime\mathrm{s}}$包含了 $\boldsymbol{I}_{\mathrm{t}}$ 和 $\boldsymbol{I}_{t\prime\mathrm{s}}$之间的所有变化,故在重建的过程中需使模型更加重视 $\boldsymbol{p}^{i}_{t\prime\mathrm{s}}$。

为了更加直观地显示相互关联元素对角线上的序列,在 $\boldsymbol{P}_{\mathrm{t}+\delta}$和 $\boldsymbol{P}_{t\prime\mathrm{s}}$之间构建了一个注意力图像 $\hat{\boldsymbol{D}}$,其中 diag($\hat{\boldsymbol{D}}$)解释了两个对应区域 $\boldsymbol{p}^{i}_{\mathrm{t}+\delta}$和 $\boldsymbol{p}^{i}_{t\prime\mathrm{s}}$之间的相关性。进行重建操作时,将得到的相关性序列与 $\boldsymbol{P}_{t\prime\mathrm{s}}$的内容向量相乘,再将这些乘积相加得到最终的输出向量,diag($\hat{\boldsymbol{D}}$)可以作为重要特征的权重。DMA 的完整操作如式(3)、式(4)所示。

$$\hat{\boldsymbol{D}} = \mathrm{Softmax}((\boldsymbol{Q}(\boldsymbol{P}_{\mathrm{t}+\delta}) \otimes \boldsymbol{K}\boldsymbol{P}_{t\prime\mathrm{s}})^{\mathrm{T}}),$$

$$\sum_{j=0}^{D_p} \hat{\boldsymbol{D}}(i,j) = 1 \tag{3}$$

$$\boldsymbol{P}_{\mathrm{dma}} = \mathrm{diag}(\hat{\boldsymbol{D}}) \times \boldsymbol{V}(\boldsymbol{P}_{t\prime\mathrm{s}}) \tag{4}$$

式中:$\boldsymbol{Q}$ 为查询向量值;$\boldsymbol{K}$ 为被查向量值;$\boldsymbol{V}$ 为内容向量值;$\hat{\boldsymbol{D}}$ 为特征权重;$\boldsymbol{P}_{\mathrm{t}+\delta}$和 $\boldsymbol{P}_{t\prime\mathrm{s}}$分别为不同图像的对应交换块;$\boldsymbol{P}_{\mathrm{dma}}$为输出的特征向量。

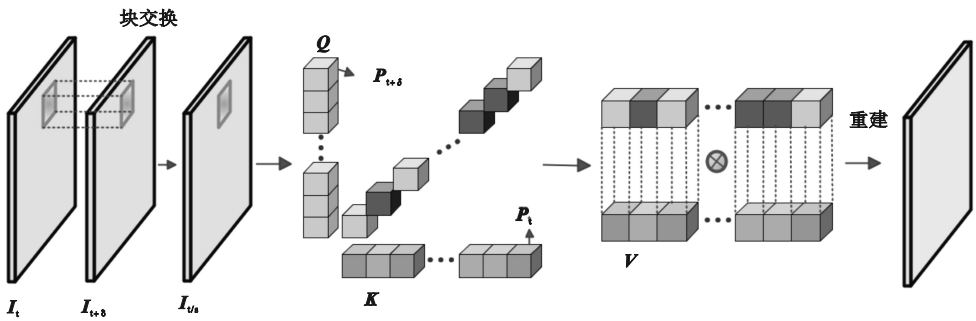

图 1 DMA 示意图

Fig. 1 Schematic diagram of DMA

2 时空差异模块

患有抑郁症的个体会表现出特定的面部表情变化模式,与时间和空间有关。自动抑郁检测的目标是为了捕获最具区分性的面部表情变化,发现面部携带的关键信息。ResNet[22] 的深层网络结构有助于模型提高表情的识别能力。该网络采用残差连接[23],允许网络学习输入到输出的残差,从而避免深层网络训练中的梯度消失问题。此外,ResNet 采用了分层结构,每一层由多个相同的残差块组成,每个残差块包含多个残差单元,使网络可以较好学习面部特征,从而提高模型的性能。本文在残差网络的基础上提出 SDN,该模块由最大化区块和差异化区块构成,用以综合时空信息与编码时空变化的细节。SDN 结构如图 2 所示。

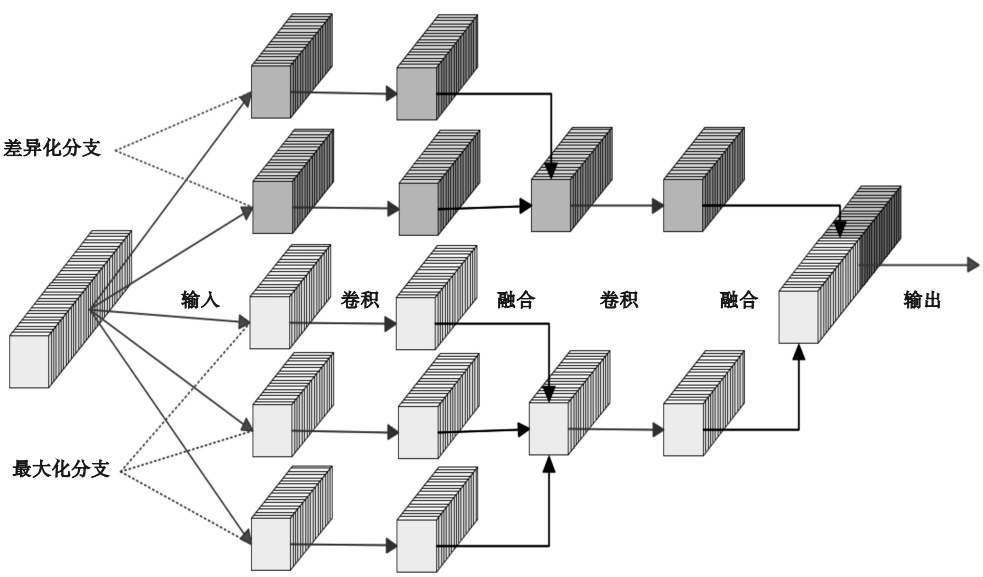

图 2 SDN 结构图

Fig. 2 Structural diagram of SDN

2.1　最大化区块

最大化区块是对全局的空间和时间变化进行建模,使用一个函数总结 2D 卷积层级联中的变化,允许模块提取相关的时空特征,进而提高模型对抑郁检测的性能。为使本文算法具备捕捉平滑面部变化的潜力,函数由一系列 3 ×3 ×3 的卷积核在视频帧上滑动并计算每个位置的最大响应值完成。设输入特征图为 $\boldsymbol{X}\in \mathrm{R}^{S\times T\times H\times W\times C}$,其中 S、T、H、W 和 C 分别为批量大小、时间深度、高度、宽度和通道数,将操作定义如下。

$$\boldsymbol{L}_{T,H,W}=\max\{\boldsymbol{X}_{T;T+Y,H,W}\} \tag{5}$$

式中:$\boldsymbol{L}_{T,H,W}$为时空的表示;Y 为用于沿深度轴执行最大池化的滑动窗口的长度。

本文提出的区块未使用固定时间深度的结构来探索时空变化,而是采用不同的动态范围帮助模型捕获对抑郁症表征的补充信息。最大化区块由 N 个分支组成,每一个分支在 $l_k(k=1,2,\cdots,N)$范围内运行,当有更多的分支时会增加参数的数量,从而增加模型的训练时间,但少量的分支又可能会降低模型的能力,故需要选择适当的分支数。设 $\boldsymbol{B}^k$ 表示分支 k 的输出,区块的输出可以表示为

$$\boldsymbol{Z}=F\{\cup_{k=1}^{N}\boldsymbol{B}^k\} \tag{6}$$

式中:$\boldsymbol{Z}$ 表示最终的特征图;$F\{\}$是由 1 ×1 ×1 卷积层进行的融合函数;∪ 是连接每个分支输出的操作。

2.2　差异化区块

当模型生成对面部变化具有鲁棒性的表示时,需要对面部结构中的突变信息进行编码,以帮助模型分析那些具有相似面部表情变化的视频片段。在 SDN 中引入差异化区块计算相邻帧之间的差异,研究面部表情变化的速度,其处理过程说明如下。

将 $\boldsymbol{X}\in \mathrm{R}^{S\times T\times H\times W\times C}$定义为输入特征图,参数设置与上文相同,然后计算特征图之间差异的绝对值 $\boldsymbol{E}_T$,计算表达式为

$$\boldsymbol{E}_T=|\boldsymbol{X}_T-\boldsymbol{X}_{T-\alpha}| \tag{7}$$

式中:α 表示第 α 阶差;$\boldsymbol{X}_T$ 是在深度 T 处的特征图;$\boldsymbol{X}_{T-\alpha}$是在深度 $T-\alpha$ 处的特征图。

与最大化区块一样,差异化区块同样由 N 个分支组成,这些分支通过执行 $\alpha_1,\alpha_2,\cdots,\alpha_N$ 阶次的差分获得时空变化的速度。这里需要确保输出的深度大小等于输入特征图的大小,通过在执行操作时将零添加到输入特征图以达到目的。由于差异化区块旨在探索短期变化,因此宜采用如 1、2 和 3 这种较低阶的差异,高阶的差异化区块更利于探索长期变化。在此过程中,生成特征图中的空间依赖性由 2D 滤波器进行探索。该区块的输出可由式(8)表示。

$$\boldsymbol{u}=F\{\cup_{k=1}^{N}\boldsymbol{G}^k\} \tag{8}$$

式中:$\boldsymbol{u}$ 表示最终的特征图;$\boldsymbol{G}^k$ 是第 k 个分支的输出。

SDN 由差异化区块和最大化区块构成,沿通道轴进行通道拼接操作,两个块的特征通过线性融合输出。使用额外的 1 ×1 ×1 卷积层来调整通道数,以使其与输入特征图匹配。本文的特征提取器基于残差网络架构,其中将 SDN 替换残差网络中除第一层之外的残差块。

3　抑郁检测流程

本文提出的基于 SDN 的抑郁检测方法结构如图 3 所示。整个抑郁检测的过程可以分为三部分:数据预处理、表情分析、得分评估。具体检测评估步骤如下。

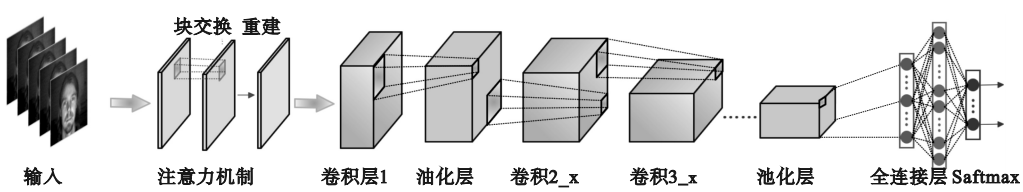

图 3　SDN 抑郁检测结构图

Fig. 3　Structure diagram of spatiotemporal difference network depression detection

1)通过采用多任务级联卷积网络(MTCNN)对所有样本进行人脸检测,然后进行人脸对齐操作,以减少姿势等问题对预测结果的影响。最后把获取的面部图像以中心区域为基准调整尺寸,以减少潜在的背景干扰。

2)通过注意力机制对输入图像进行块交换等操作,增强模型识别与发现图像间变化的能力。再利用 SDN 对面部表情进行捕捉与分析。

3)通过在网络中得到的数据转化为 BDI-II 得分,对患者进行抑郁程度评估。

4 实验验证

4.1 数据集

为验证本文方法的有效性,采用 ACM 国际多媒体会议期间举办的视听情感挑战和研讨会所建立的数据集,即 AVEC 2013[24] 和 AVEC 2014[25] 视听抑郁数据集。

AVEC 2013 数据集源自视听抑郁语言语料库(AViD-Corpus)的子集。受试者在完成计算机上安排的交互任务时相关数据被记录下来,该数据集共包含 150 个视频剪辑,每一个都统一分为训练集、验证集和测试集三个子集。每组由 50 个不同样本的视频组成,每个视频都有与受试者抑郁评分相关的标签。这些视频的时长约为 20 ~ 50 min,平均视频时长为 25 min。

AVEC 2014 数据集也是 AViD-Corpus 的子集,是专门为情绪分析和抑郁症识别等领域的研究而制定的。该数据集以录制视频的形式,记录了受试者执行两个名为 Freeform 和 Northwind 的交互任务。在 Freeform 任务中,受试者需要回答比如描述出悲伤的童年记忆等问题。在 Northwind 任务中,受试者需要大声朗读寓言摘录。在这两个任务中,视频都被统一分为训练集、验证集和测试集三个子集。每组包含 50 个视频,每个视频都有其对应的标签。该数据集共由 300 个视频组成,视频时长在 6 ~ 248 s 之间。两个数据集中视频的帧速率均为 30 s^{-1}。

本文的目标是评估样本的 BDI-II 得分:轻微或无(0 – 13)、轻度(14 – 19)、中度(20 – 28)和重度(29 – 63)。为对所提出结构的性能进行评估,以及与其它抑郁识别方法进行公平比较,本文采用平均绝对误差(MAE)和均方根误差(RMSE)两个指标。

4.2 实验结果与分析

在 AVEC 2013 和 AVEC 2014 中,本文从视频的每 10 帧中选择 1 帧作为模型的输入,把获取的面部图像以中心区域为基准将尺寸调整为 224 × 224,批量大小设为 64。为了优化模型,选择权重衰减系数为 0.000 1 的 Adam 优化器,学习率设为 0.000 2,用于分类的最后一个全连接层学习率设为 0.002。使用 Gamma 为 0.9 的 ExponentialLR 学习率调度器,以降低每个训练轮次后的学习率。

为了评估模型的能力,首先对 SDN 使用 FER 2013 和 RAF-DB 数据集进行预训练,使其具备探索面部结构的能力。使用 AVEC 2013 和 AVEC 2014 对其微调测试,性能如表 1 所示。由表 1 可见,当使用大型数据集对 SDN 进行预训练时,可以有更好的性能。模型在 RAF-DB 上拥有最佳的性能,可能是因为 FER 2013 数据集存在标签缺失、标签错误等问题,而且图片的分辨率也较低。

表 1 使用不同数据集预训练性能分析

Table 1 Performance analysis of pre-training using different datasets

预训练	AVEC 2013		AVEC 2014	
	RMSE	MAE	RMSE	MAE
无训练	9.40	7.56	9.09	7.39
FER 2013	9.12	7.22	8.76	6.95
RAF-DB	**9.03**	**7.08**	**8.66**	**6.95**

接下来需要研究 SDN 中最大化区块和差异化区块的分支数量变化所带来的影响,因为需要训练多个模型,训练时间较长,所以仅使用 50 层的 SDN 在 AVEC 2014 数据集上进行分析,结果如表 2 所示。在每一层中深度值是最大化区块探索时空变化的长度,阶数是指差异化区块的阶数,值代表分支的数量。从表 2 中可以看出,当同时使用最大化区块和差异化区块时,比仅使用其中一个区块时的结果更好,反映出探索面部平滑和突然变化信息的重要性。综合来看,采用三个分支的最大化区块和两个分支的差异化区块在 RMSE 和 MAE 方面都有比较出色的表现。基于上述结果,本文使用三分支的最大化区块和两分支的差异化区块来构成 SDN。

因整体框架中的 SDN 是基于 3D-ResNet 网络进行改进的,所以首先在 RMSE、MAE 和计算复杂度方面将 DSN 方法与 3D-ResNet 进行比较。性能分析结果如表 3 所示,表中所提模型已按照与本文提出方法相同的程序进行训练。

从表 3 中的结果可以看出,模型的性能随着网络深度的增加而逐渐提高。与规模较小的 DSN-18 相比,DSN-100 与 DSN-152 已经实现了较大的性能改进。与 3D-ResNet 相比较,DSN 在 RMSE 和 MAE 方面均表现出更好的性能,而且参数量与运算量也要小得多。3D-ResNet 在网络较深时需要大量的计算资源来训练和推理,且可能会过拟合,需要通过正则化等方法进行处理。由此可见,DSN架构要比3D-ResNet更适合抑郁检

表 2　对具有不同模块配置的 SDN-50 的评估

Table 2　Evaluation of SDN-50 with different module configurations

层								AVEC 2014	
Res2-x		Res3-x		Res4-x		Res5-x			
深度	阶数	深度	阶数	深度	阶数	深度	阶数	RMSE	MAE
4		3		2		1		9.44	7.85
	1		1		1		1	9.52	7.96
4	1	3	1	2	1	1	1	8.98	7.10
3	1	2	1	1	1	1	1	9.32	7.48
2	1	1	1	1	1	1	1	9.64	7.10
4	1,2	3	1,2	2	1,2	1	1,2	8.64	6.70
3	1,2	2	1,2	1	1,2	1	1,2	9.36	7.37
2	1,2	1	1,2	1	1,2	1	1,2	9.20	7.08
3,4	1	2,3	1	1,2	1	1,1	1	9.00	6.92
2,3	1	1,2	1	1,1	1	1,1	1	8.75	6.71
2,3	1,2	1,2	1,2	1,1	1,2	1,1	1,2	8.40	6.53
3,4	1,2	2,3	1,2	1,2	1,2	1,1	1,2	8.37	6.58
2,3	1,2,3	1,2	1,2,3	1,1	1,2,3	1,1	1,2,3	8.35	**6.41**
2,3,4	1,2	1,2,3	1,2	1,1,2	1,2	1,1,1	1,2	**8.16**	6.45

表 3　DSN 和 3D-ResNet 的性能分析

Table 3　Performance analysis of DSN and 3D-ResNet

网络	AVEC 2013		AVEC 2014		参数量/10^6	浮点运算量/10^9
	RMSE	MAE	RMSE	MAE		
ResNet-18	9.24	7.06	9.14	6.92	33	8.38
DSN-18	8.96	7.21	8.82	6.77	7	5.66
ResNet-50	8.81	6.92	8.40	6.79	63	12.22
DSN-50	8.13	6.39	8.16	6.45	21	7.40
ResNet-101	8.51	6.79	8.20	6.57	121	17.80
DSN-100	7.62	**6.14**	7.92	6.21	36	10.34
ResNet-152	8.30	6.58	8.01	6.30	168	24.70
DSN-152	**7.55**	6.24	**7.65**	**6.06**	52	13.36

测,可以更准确地克服模糊性和过拟合等问题,尤其是对于网络规模较大的模型。

将 DSN 与 Inflated 3D(I3D)[26] 和 Temporal 3D(T3D)[27] 模型进行比较,性能分析结果如表 4 所示。I3D 模型由一个称为初始模块的基本结构组成,该结构通过膨胀 2D 滤波器和池化 2D 版本模块的内核而获得,可以捕捉视频中的时空信息。T3D 模型则包含时间过渡层的结构,负责捕获不同范围内的时间信息。两种架构已成功应用于行为识别,与此类模型的比较对于衡量 DSN 架构的功能性较为重要。

由表 4 可知,与 T3D 相比,I3D 和 DSN 使用更少的参数量就可以得到更好的性能,且 DSN 的 RMSE 值显示出其性能的优势。在计算复杂度方面,T3D 的运算量大,I3D 和 DSN 以更低的运算量取得了更好的性能效果。虽然 I3D 比 DSN 使用的参数更少,但其性能有较大差距,因为 DSN 旨在探索面部平滑和突然变化的时间信息。综合来看,基于互补功能和不同深度的 DSN 结构在捕捉面部表情的时空变化方面有良好的潜力,即使在有限的训练数据下,该架构也可以学习如何获得面部表情变化的丰富表示。因此,该模型在探索用于抑郁症检测的面部视频方面具有良好的性能。

表 4 DSN、I3D 与 T3D 架构性能分析

Table 4 Performance analysis of DSN, I3D and T3D architectures

网络	AVEC 2013		AVEC 2014		参数量/10^6	浮点运算量/10^9
	RMSE	MAE	RMSE	MAE		
I3D	8.66	6.64	8.55	6.36	13	6.99
T3D	8.75	6.76	8.55	6.54	68	51.64
DSN-50	8.13	6.39	8.16	6.45	21	7.40
DSN-100	7.62	**6.14**	7.92	6.21	36	10.34
DSN-152	**7.55**	6.24	**7.65**	**6.06**	52	13.36

图 4 为本文方法 DSN-152 分别在 AVEC 2013 和 AVEC 2014 上的误差可视化图,其中呈现的是由最小误差率到最大误差率的视频排序。由图 4 可知,误差概率在 0 ~ 13 之间近似均匀分布,只有少数的值超过该区间。在 AVEC 2013 和 AVEC 2014 中,超过 60% 的样本模型误差小于 6,6 左右的误差表示对抑郁症严重程度的错误分类仅在相邻类别之间和类别边界处的分数。当误差大于 16 时,表示抑郁程度最低的样本被归类为重度抑郁,该情况的原因为个体之间存在差异,在模型标准化评估下,同样的面部表情在不同个体中存在不同的含义。本文提出的方法在 AVEC 2013 和 AVEC 2014 中只有 6 个样本的误差大于 16,这些结果表明,DSN-152 具有良好的泛化能力,重大错误分类的可能性很小。

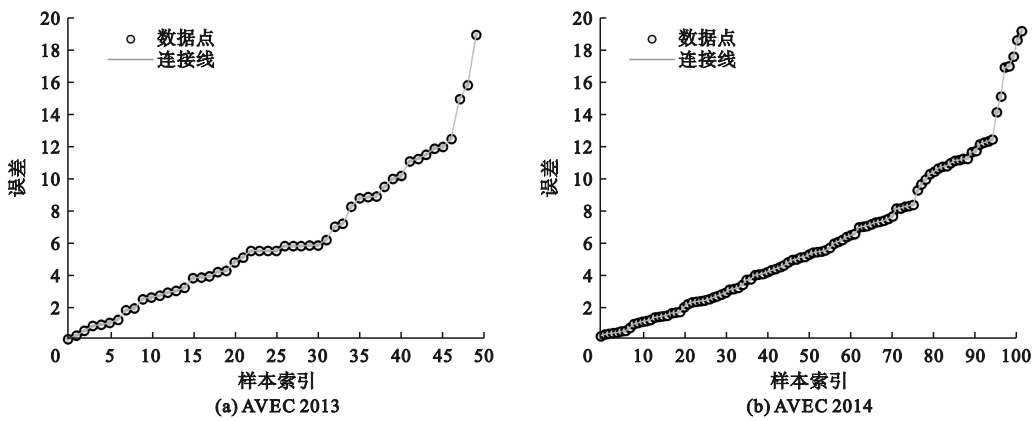

图 4 在 AVEC 2013 和 AVEC 2014 的误差可视化图

Fig. 4 Visualization of errors in AVEC 2013 and AVEC 2014

5 结论

本文提出了一种基于时空差异网络的抑郁检测方法 DSN,并在用于人脸视频抑郁症检测的基准数据集(AVEC 2013 和 AVEC 2014)上对该方法的性能进行了评估,得到的结论如下。

1)DSN 的 DMA 机制通过以较低的计算量使模型关注面部重要区域,从而提升模型的性能,减轻计算负担。

2)DSN 可以在不依赖 3D 卷积的情况下捕获多范围特征,不仅保持了模型性能的稳定,而且显著降低了运算复杂度,进一步减小了过拟合的风险。

3)DSN 在有限的训练数据下,能够学习获取丰富的面部表情变化表示,展现出良好的鲁棒性和泛化性。

参考文献(References):

[1] BELMAKER R H, AGAM G. Major depressive disorder[J]. New England Journal of Medicine, 2008, 358(1): 55 – 68.

[2] OTTE C, GOLD S M, PENNINX B W, et al. Major depressive disorder[J]. Nature Reviews Disease Primers, 2016, 2: 16065.

[3] ACHARYA U R, OH S L, HAGIWARA Y, et al. Automated EEG-based screening of depression using deep convolutional neural network[J]. Computer Methods and Programs in Biomedicine, 2018, 161: 103 – 113.

[4] WU H F, LU B Q, ZHANG Y, et al. Differences in prefrontal cortex activation in Chinese college students with different severities of depressive symptoms: a large sample of functional near-infrared spectroscopy (fNIRS) findings[J]. Journal of

Affective Disorders,2024,350:521 – 530.
[5] SHANGGUAN Z X,LIU Z Y,LI G,et al. Dual-stream multiple instance learning for depression detection with facial expression videos[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering: a Publication of the IEEE Engineering in Medicine and Biology Society, 2023, 31: 554 – 563.
[6] ZHOU X Z,HUANG P,LIU H M,et al. Learning content-adaptive feature pooling for facial depression recognition in videos[J]. Electronics Letters,2019,55(11):648 – 650.
[7] WEN L Y,LI X,GUO G D,et al. Automated depression diagnosis based on facial dynamic analysis and sparse coding[J]. IEEE Transactions on Information Forensics and Security, 2015,10(7):1432 – 1441.
[8] JAN A,MENG H Y,GAUS Y F B A,et al. Artificial intelligent system for automatic depression level analysis through visual and vocal expressions[J]. IEEE Transactions on Cognitive and Developmental Systems,2018,10(3):668 – 680.
[9] ZHOU X Z,JIN K,SHANG Y Y,et al. Visually interpretable representation learning for depression recognition from facial images[J]. IEEE Transactions on Affective Computing, 2020,11(3):542 – 552.
[10] ZHOU X Z,WEI Z Q,XU M,et al. Facial depression recognition by deep joint label distribution and metric learning[J]. IEEE Transactions on Affective Computing, 2022, 13(3): 1605 – 1618.
[11] HE L,GUO C G,TIWARI P,et al. Intelligent system for depression scale estimation with facial expressions and case study in industrial intelligence[J]. International Journal of Intelligent Systems,2022,37(12):10140 – 10156.
[12] NIU M Y,TAO J H,LIU B. Multi-scale and multi-region facial discriminative representation for automatic depression level prediction[C]//IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP). Toronto,Canada:IEEE,2021:1325 – 1329.
[13] NIU M Y,HE L,LI Y,et al. Depressioner: facial dynamic representation for automatic depression level prediction[J]. Expert Systems with Applications,2022,204:117512.
[14] JAZAERY M A,GUO G D. Video-based depression level analysis by encoding deep spatiotemporal features[J]. IEEE Transactions on Affective Computing, 2021, 12(1): 262 – 268.
[15] ZHU Y,SHANG Y Y,SHAO Z H,et al. Automated depression diagnosis based on deep networks to encode facial appearance and dynamics[J]. IEEE Transactions on Affective Computing,2018,9(4):578 – 584.
[16] YU J H,GAO H W,CHEN Y Q,et al. Deep object detector with attentional spatiotemporal LSTM for space human-robot interaction[J]. IEEE Transactions on Human-Machine Systems,2022,52(4):784 – 793.
[17] YU J H,GAO H W,CHEN Y Q,et al. Adaptive spatiotemporal representation learning for skeleton-based human action recognition[J]. IEEE Transactions on Cognitive and Developmental Systems,2022,14(4):1654 – 1665.
[18] YU J H,GAO H W,ZHOU D L,et al. Deep temporal model-based identity-aware hand detection for space human-robot interaction[J]. IEEE Transactions on Cybernetics, 2022, 52(12):13738 – 13751.
[19] YU J H,XU Y K,CHEN H,et al. Versatile graph neural networks toward intuitive human activity understanding[J]. IEEE Transactions on Neural Networks and Learning Systems,2022,35(7):8869 – 8881.
[20] YU J H, GAO H W, SUN J, et al. Spatial cognition-driven deep learning for car detection in unmanned aerial vehicle imagery[J]. IEEE Transactions on Cognitive and Developmental Systems,2022,14(4):1574 – 1583.
[21] 于耀淋,张景异,睢付佳. 基于生成式对抗网络的人脸图像生成[J]. 沈阳理工大学学报,2022,41(5):29 – 33.
YU Y L,ZHANG J Y,JU F J. Face image generation based on generative adversarial network[J]. Journal of Shenyang Ligong University,2022,41(5):29 – 33. (in Chinese)
[22] HU F X,LIU T L,TAO D C. Why ResNet works? Residuals generalize[J]. IEEE Transactions on Neural Networks and Learning Systems,2020,31(12):5349 – 5362.
[23] 付丽君,张齐鹏,姜宇宏,等. SSA 和 DSC-ResNet 的 TE 过程故障诊断方法[J]. 沈阳理工大学学报,2021,40(3):14 – 18.
FU L J, ZHANG Q P, JIANG Y H, et al. Fault diagnosis method of TE process based on SSA and DSC-ResNet[J]. Journal of Shenyang Ligong University, 2021, 40(3): 14 – 18. (in Chinese)
[24] VALSTAR M,SCHULLER B,SMITH K,et al. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge[C]//Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge. Barcelona, Spain:ACM,2013:3 – 10.
[25] VALSTAR M,SCHULLER B,SMITH K,et al. AVEC 2014: 3D dimensional affect and depression recognition challenge [C]//Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. Orlando, USA:ACM,2014:3 – 10.
[26] HUANG Y K,GUO Y C,GAO C. Efficient parallel inflated 3D convolution architecture for action recognition[J]. IEEE Access,2020,8:45753 – 45765.
[27] GUO S N,LIN Y F,LI S J,et al. Deep spatial-temporal 3D convolutional neural networks for traffic data forecasting[J]. IEEE Transactions on Intelligent Transportation Systems, 2019,20(10):3913 – 3926.

(责任编辑:和晓军)