基于3D-CNN和时空注意力-卷积LSTM的抑郁症识别研究*

何 浪[1,2]

(1.西安邮电大学计算机学院,陕西 西安 710121;2.西安邮电大学陕西省网络数据分析与智能处理重点实验室,陕西 西安 710121)

摘要:针对目前抑郁症识别方法中没有有效地从序列信息中刻画长时特征表示的问题,本文提出了一种基于三维卷积神经网络(3D-CNN)、时空注意力-卷积长短时记忆网络(STA-ConvLSTM)和空间金字塔池化(SPP)的端到端的抑郁症识别方法,计算在数据库AVEC2013和AVEC2014测试集上的抑郁症识别结果,并将本文结果与已发表文献结果进行对比分析.结果表明:在数据库AVEC2013和AVEC2014测试集上,平均绝对误差均为8.60,均方根误差分别为10.32和10.27,文献结果比较表明本文的抑郁症识别结果优于当前大多数基于面部视频的抑郁症识别方法.

关键词:抑郁症;时空注意力机制ConvLSTM(STA-ConvLSTM);三维卷积神经网络;卷积长短时记忆神经网络

中图分类号:TP391.4 **DOI**:10.19789/j.1004-9398.2021.02.004

0 引 言

近年来,情感计算领域采用信号处理、计算机视觉和机器学习等方法来分析抑郁症患者的言语行为和非言语行为.在研究者认为:面部非言语行为的动态激活,对抑郁症的自动估计起着非常重要的作用,而且至少有一半的非言语行为都集中在面部进行研究[1].根据观察到的非言语行为,基于头部姿势、面部表情和活动等,已提出了大量的抑郁症分析系统和方法[2-8].目前,基于听觉抑郁症识别和分析方法还存在以下问题:(1)动态序列特征的模式信息还没有得到有效的挖掘,其能够有助于更准确地预测抑郁症患者的抑郁程度,提升抑郁症的识别精度;(2)传统的抑郁症识别方法通常包括特征提取、特征聚集和回归(或分类)3个步骤,这些步骤可逐步串联后对抑郁症进行评估,如何将这3个步骤集成为一个端到端的方法进行抑郁症的估计,对于提高抑郁症识别的精度是至关重要的.

1 相关工作

目前,随着计算机视觉和人工智能的迅速发展,基于视觉抑郁症的识别方法越来越得到国内外研究者的重视.本文进行了国内外文献调研,国外研究显示,抑郁症研究由原来的手工设计特征逐步转移到深度学习特征进行特征抽象,国内研究显示,抑郁症研究也大多数集中于采用深度学习方法进行特征的提取.以下分别进行总结.

在国外,在手工设计特征方面,目前常用的是局部伽柏二值模式-三维正交平面[12-13]和局部相位量化-三维正交平面[3]等,这些特征难以获取准确的、全局的细微面部表情信息,而且部分特征维数高、计算效率低,非常不利于构建自动化的抑郁症识别系统或者框架.Al和Guo[6]提出了一种从视频数据中估计BDI-II分数的新框架,该框架根据三维卷积神经网络(3D convolutional neural network,C3D)自动学习人脸区域2种不同尺度的时空特征,采用递归神经网络(recurrent neural network,RNN)对时空信息序列进行进一步的学习表征,通过在AVEC2013和AVEC2014上进行实验,该方法取得优于当前大多数方法的识别结果;Song等[7]将人类行为原语作为每帧的底层特征进行特征提取,在视频级上,通过谱特征和谱向量来表征多尺度面部表

收稿日期:2020-06-17

*陕西省教育厅科研计划项目(20JG030);陕西省普通高等学校重点学科专项资金建设项目;陕西省高等教育学会"疫情防控专项研究课题"(XGH20201);陕西省提升公众科学素质计划(新型冠状病毒感染肺炎应急科普)(2020PSL(Y)040)

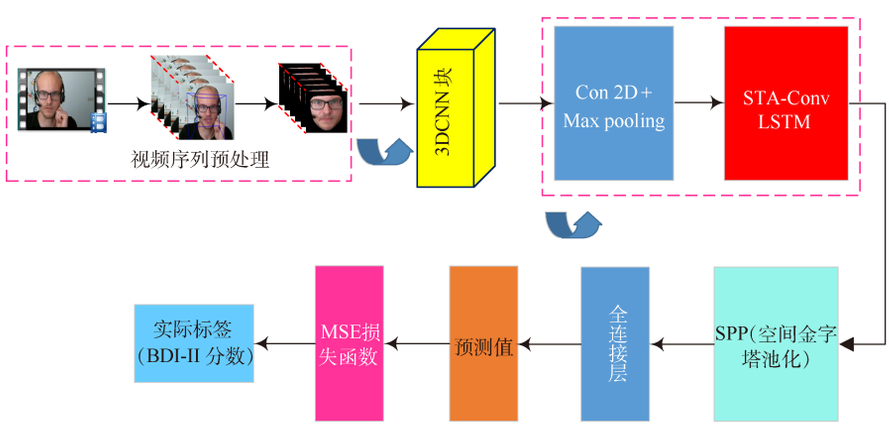
情模式信息,使用卷积神经网络和人工神经网络进行抑郁症的估计,在AVEC2013和AVEC2014上对该方法进行验证,得到了优于大多数方法的识别效果;Azher等[8]提出了一种新的抑郁症估计框架,利用双流深度时空特征从视频数据中估计抑郁尺度,通过Inception ResNetv2网络进行微调来提取空间信息,同时引入了一个本地卷方向数的动态特征描述符来捕捉面部运动,将本地卷方向数的动态特征描述符输入卷积神经网络以获得更多的鉴别特征,利用一个多层双向长短时记忆网络(BiLSTM)模型,基于聚集序列中位数的方法获取了整个视频的全局特征表征,并在AVEC2013和AVEC2014上进行验证,得到目前基于视频模态的最好结果.

在国内,Chao等[15]将语音的手工设计特征、面部图像作为输入,采用LSTM进行回归预测,在AVEC2014上进行验证,取得了优于大多数听视觉多模态融合的抑郁症识别结果;Zhou等[5]提出深度回归网络DepressNet来学习一种带有视觉理解的抑郁特征表示,首先采用CASIA大型人脸数据库上预训练一个带有全局平均池层深度卷积神经网络ResNet[16],然后在AVEC2013和AVEC2014上进行微调,微调以后的DepressNet可以根据单幅图像的抑郁症分数,生成对应的抑郁特征激活图来识别输入图像的显着性区域,结果表明,该方法显著地提高了基于视觉的抑郁症识别的准确率,而且眼睛附近的区域对于抑郁症的识别起着至关重要的作用;Yang等[17]提出了深层卷积神经网络深层神经网络(deep convolutional neural network-deep neural network,DCNN-DNN)的框架,首先采用openSMILE工具包提取每一个样本的全局特征向量,然后输入到DCNN-DNN深层模型中进行二次特征抽象,最后采用随机森林树对正常人和抑郁症病人进行分类,取得了优于大多数方法的分类结果.

2　基于3D-CNN和卷积LSTM的抑郁症识别方法

借鉴于深度学习的特征表征能力,为了充分表征蕴含在上下文的局部和全局时空序列信息,本文提出了基于三维卷积神经网络(3D convolutinal neural network,3D-CNN)和时空注意力-卷积长短时记忆网络(spatio-temporal attention based ConvLSTM,STA-ConvLSTM)的抑郁症识别方法.该方法首先采用3D-CNN网络来学习短时的局部时空特征,然后采用STA-ConvLSTM来抽象和表征长时的、具有注意力模式信息的全局特征表征,最后使用空间金字塔池化(spatial pyramid pooling,SPP)进行不同的特征尺度变换[9].选择3D-CNN的原因主要是因为其具有学习和表征视频序列的能力,在特征表示过程中避免了时域的重叠.STA-ConvLSTM模块的优势在于:(1)学习一种基于时间方向的上下文注意力表示;(2)表征和建模最相关的时间动态信息.上下文的时空注意力信息通过ConvLSTM进行进一步的学习[10],能够挖掘整个时间序列的显式和隐式内容.

基于3D-CNN、STA-ConvLSTM、SPP的抑郁症识别方法如图1所示.该方法包括:(1)使用开源框架OpenFace进行人脸检测和对齐[18];(2)使用3层的3D-CNN来抽象和表征短时的时空序列特征,为了避免过拟合和信息冗余,接着采用2D卷积和最大池化进行降采样,实现局部短时时空信息的有效表征;(3)为了从这些局部时空特征中学习和抽象全



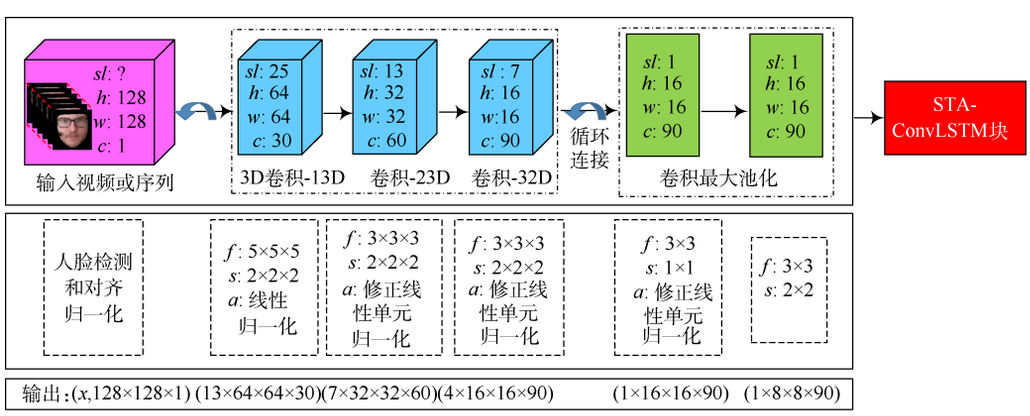
注:3D-CNN为三维卷积神经网络;STA-ConvLSTM为时空注意力-卷积长短时记忆网络;SPP为空间金字塔.

图1　基于3D-CNN和STA-ConvLSTM的抑郁症识别方法

局时空特征,对全局结构信息进行编码,采用STA-ConvLSTM能够有效地进行时空序列信息的中间抽象,从而得到一个只关注整个输入序列中相关特征的高层语义表示;4)采用SPP进行特征的变换,利用不同的参数在固定的子区域上聚集多尺度的局部特征,得到紧致的最终特征表征;(5)通过采用MSE损失函数,得到BDI-II分数的预测.

2.1　局部时空序列特征提取

一般来说,从图像序列中直接提取局部时空特征,不仅能够获取特征的空间信息,而且考虑了动态信息的连续性.在本文中,假设全局时空特征是由局部时空和运动特征按照一定的尺度来表征.因此,设计了一种3D-CNN的体系结构,能够从输入的面部帧生成许多个信息通道,通过进行不同尺度的三维卷积,从而捕获在多个相邻帧中编码的动态信息.基于3D卷积神经网络的特征提取能力,设计了如图2所示的3D卷积结构用于局部时空序列特征的提取.将面部图像裁剪为128×128像素,采用灰度通道作为3D-CNN的输入.为了更好扩展和应用于其他任务,视频的序列长度是可变的.在任务中,采用25帧作为视频子序列的长度,然后将最终的输出输入到STA-ConvLSTM模块中,在连续循环的过程中,序列中的相邻帧没有重叠.为了抽象时空依赖信息和整个视频序列的语义、上下文信息,3次循环的权重是彼此共享的.3D卷积块由3层卷积级联组成,第1层采用线性激活函数,滤波器大小为5×5×5.第1个卷积层的维度为25×64×64×30,其中第一维表示时间信息(sl=25),第二维(h=64)和第三维(w=64)表示空间信息,第四维(c=30)表示通道信息.为了降低时间和空间的维度,这一层的步长设置为2,滤波器的个数为30个.为了避免过拟合和降低信息冗余,每一层都要进行丢包操作.从本质上说,丢包操作意味着以一定的概率随机丢弃层里面一定数量的神经元(以及他们之间的连接).为了能够加快模型训练的收敛,每一层都要进行归一化.对于第2和3层,采用相同的参数进行设置,区别在于滤波器大小改变为3×3×3,特征图数目分别为60和90个.另外,ReLU函数被用作激活函数.如果输入时间步长k_1是25×128×128×1,通过第1个3D卷积之后,第3个3D卷积层的输出是4×16×16×90;将其输入到2D卷积层进行下采样,由此可以得到4×8×8×90的输出,然后采用STA-ConvLSTM块进行全局依赖信息的抽象和变换.通过级联使用3D卷积块,对连续帧之间的局部变化进行刻画和编码,能够有效处理很长的输入序列,得到紧致的特征表征.通过使用3D和2D卷积及最大池化,可以对短时时空依赖信息进行深层次的编码和刻画,但是在这些特征信息中间还有部分的冗余,需要进一步的抽象和表征.因此,本文提出了STA-ConvLSTM,采用软性注意力机制进行进一步的特征学习和表征.



注:sl为序列长度;h为高度;w为宽度;c为通道;f,s,a是可变的.

图2　表征局部变化的3D卷积块

2.2　全局时空序列特征提取

对于听视觉抑郁症识别和分析来说,音视频信号中含有众多缓慢变化的、长时的特征信息,需要进行不同阶段的中间处理和抽象,然后才能得到对分析抑郁症至关重要的信息.因此,提出空间软性注意力和时序软性注意力模型STA-ConvLSTM的全局特征抽象方法,该方法不仅可以避免时序信息的丢失,而且可以抽象和表征整个序列全局语义信息

的高层表示.

本文提出STA-ConvLSTM进行全局时空依赖信息的抽象和表征,STA-ConvLSTM的结构如图3所示.每个输入X_t的维度为1×8×8×90,分别表示时间步长、空间维数和通道数.每个ConvLSTM单元的滤波器大小为3×3,含有45个特征图.STA-ConvLSTM实际上是由3个ConvLSTM组成,其Dropout率为65%.为了加快模型训练的收敛,对空间注意力模块进行了归一化操作.空间注意力模块和时间注意力模块的输出维度均为1×8×8×45,最后的输出仅考虑最后的一个时间步,因为其是从整个序列中获取的信息的积累,这一阶段的输出维度为1×8×8×90.

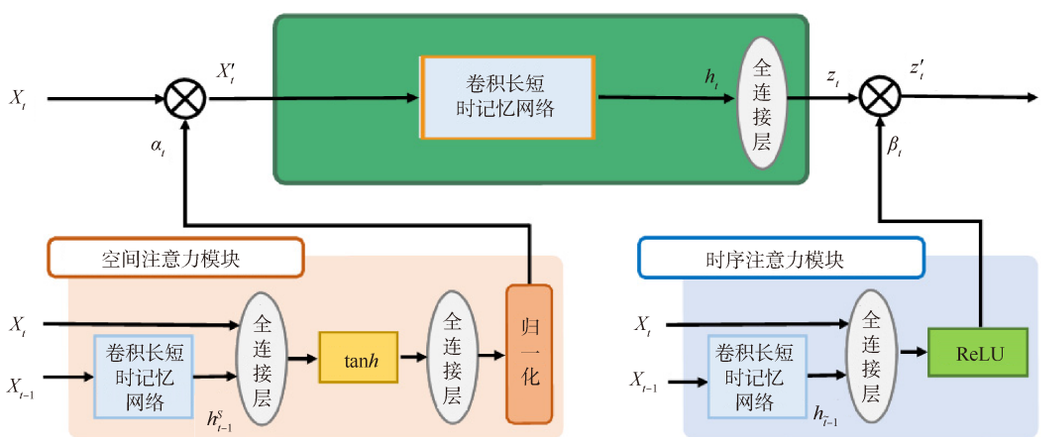


图3　STA-ConvLSTM结构

在每个时间步t,给定输入维度大小为K的$\boldsymbol{x}_t=(\boldsymbol{x}_{t,1},\boldsymbol{x}_{t,2},\cdots,\boldsymbol{x}_{t,K})^{\mathrm{T}}$,其中$\boldsymbol{x}_{t,K}\in\mathbf{R}$,权值分数$\boldsymbol{s}_t=(s_{t,1},\cdots,s_{t,K})^{\mathrm{T}}$可以估计空间维度上的重要性,即$\boldsymbol{s}_t=U_s\tanh(W_{xs}*\boldsymbol{x}_t+W_{hs}*\boldsymbol{h}^s_{t-1}+\boldsymbol{b}_s)+\boldsymbol{b}_{us}$其中,$\boldsymbol{s}_t$也叫做注意力打分函数,$U_s$、$W_{xs}$和$W_{hs}$是通过学习到的参数,$\boldsymbol{b}_s$、$\boldsymbol{b}_{us}$分别是偏置向量.$\boldsymbol{h}^s_{t-1}$是一个ConvLSTM隐藏层的变量.对于选择的第$k$个空间特征的概率$\alpha_{t,k}$为

$$\alpha_{t,k}=\frac{\exp(s_{t,k})}{\sum_{i=1}^{K}\exp(s_{t,i})},\tag{1}$$

式中$\alpha_{t,k}$是一个归一化的注意力权重分数.一系列的权重参数控制了流向主要ConvLSTM层的信息数量.在不同的空间信息中,分数越大,代表选择的信息就越重要,为后续的进一步处理提供了强有力的特征表征.特别强调的是,空间注意力机制根据当前时间步长的所有空间特征和ConvLSTM层的隐藏变量来确定特征的重要性.一方面,隐藏层变量$\boldsymbol{h}_t$包含了过去的累积信息,能够累积长时的动态信息;另一方面,为确定目前时间步内的所有特征的重要性提供了必要的保障.

2.2.1　空间注意力

经由2D卷积和池化之后的4D张量维度为1×8×8×90,通过采用软性注意力机制,空间维上被分配不同的权重,这使模型能够自适应地将更多的注意力集中在区分性的特征上,进行抑郁症的识别和分析.如图3所示,空间注意力模块由1个ConvLSTM层、2个全连接层和一个归一化单元组成.

2.2.2　时间注意力

对于时间序列,不同时间维度含有的信息数量通常是不相同的.只有一些关键的时间维包含具有鉴别性的信息,而其他维度提供上下文信息.在此基础上,设计了一个时间注意力模块,对不同的时间维特征进行不同的权重计算,自动进行不同程度的注意力表征.

$$\beta_t=\mathrm{Re}LU(\boldsymbol{w}_{x\sim}*\boldsymbol{x}_t+\boldsymbol{w}_{h\sim}*\boldsymbol{h}^{\sim}_{t-1}+b_{\sim}),\tag{2}$$

式中$\boldsymbol{h}^{\sim}_{t-1}$是时间$t$-1时ConvLSTM单元的隐藏层的输出.如图5所示,时间注意力模块由ConvLSTM层、全连接层和ReLU函数组成.与空间注意力模块相同,β_t是一个软性注意力机制的权重分数.

2.3　时空特征变换表征

为了提高识别的准确率,通过在当前的深层网络中寻找自然的特征表征信息,在空间和时间上捕捉到独特的深度面部特征.由于3D卷积、2D卷积以及STA-ConvLSTM层的权重共享,本文直接处理了面部图像位置和角度的变化.然而,由于与摄像机的距离不同和头部姿势的影响,图像序列中的人脸

尺寸有一定的变化,头部姿态变化可能导致模糊效应,影响图像的分辨率.

为了进行尺度不变的特征表示,在STA-ConvLSTM层输出的基础上,利用SPP(如图4所示)层进行多尺度特征提取.通过3D-CNN进行局部短时特征的提取和STA-ConvLSTM进行全局长时信息的进一步抽象,得到的输出维度为1×8×8×90,如图6所示,在不同的尺度上分别使用了不同的窗大小进行特征的表征,即2×2、4×4、8×8,得到的输出分别为1 440、360和90维,原始的特征直接展开为5 760维,将所有的特征向量串联,得到输出向量为7 650维.SPP层是空间金字塔匹配的扩展,可将其中任意区域(子图像)中的特征池化到一起,将图像划分为更加精细的层次,有效聚集重要特性.SPP层允许增加尺度不变性,并能够降低一定的过拟合.本文采用了文献中定义的空间金字塔核的原始加权方法,并将其与SPP相关联,其中高分辨率的特征图比低分辨率的特征图具有更高的权重,其时空特征的权重定义为

$$\frac{1}{2^L}M^0+\sum_{l=1}^{L}\frac{1}{2^{L-l+1}}M^l, \tag{3}$$

式中L是全部的尺度数目,l是当前尺度;M^l是初始化的高分辨率特征图;M^0是低分辨率的特征图.

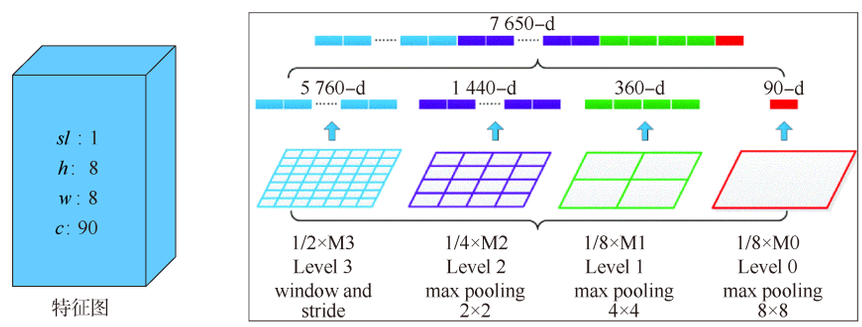


图4　SPP层结构

通过SPP层的特征尺度变换,得到7 650维的加权时空特征向量.经过前面几步抽象得到的特征,然后通过一个全连接层进行特征的再次变换.与STA-ConvLSTM抽象得到的特征相比较,经过SPP变换后的特征表征了不同尺度的模式信息.

3　抑郁程度估计

本文提出的3D-CNN和STA-ConvLSTM的抑郁程度估计框架如图5所示,通过采用3D-CNN进行局部时空特征提取,采用2D卷积进行下采样,STA-ConvLSTM进行全局时空特征提取,SPP进行特征的尺度不变表示,并应用于最终抑郁程度的估计.

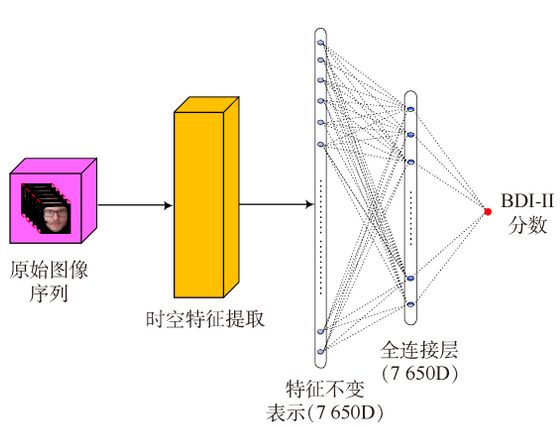


图5　抑郁程度估计

4　实验设置

4.1　数据预处理

对于AVEC2013和AVEC2014数据库,先对视频序列进行预处理,进行人脸检测和对齐.对齐的面部图像被裁剪并调整为128×128像素.由于AVEC2013和AVEC2014数据库中实验者的头部摆动的幅度较大,或有的实验者用手捂住面部、用手托着下巴和头发覆盖大半个面部等,为了提升最终的预测结果,需要手工检查检测到的人脸图像.通过人工对每一帧人脸图像进行检查,如果在人脸图像中没有检测到人脸,就删除对应的帧.在AVEC2014数据库中,99.14%的帧包含人脸图像,而在AVEC2013数据库中,只有95.61%的帧检测到人脸,这是由于数据库录制的环境和受试者的头部摆动幅度太大所造成的.同时,由于定义了输入模型的视频子序列长度为25帧,因此,将所有原始视频的帧数都切割为25帧的倍数.即将原始视频切割为长度为25帧的子视频序列,采用原始视频的帧

数除以25,如果剩余的帧数>12帧,则将剩余的帧数通过复制前一帧的方式补为25帧,如果剩余的帧数<12帧,则丢弃原始视频剩余的帧数.由于AVEC2013数据库只有1个任务,而AVEC2014分为Freeform和Northwind 2个子任务.因此,将Freeform和Northwind子任务进行合并,然后作为一个数据集进行预测,这样不仅增大了数据的样本量,而且增加了样本的多样性,同时也融合了2种不同任务的优点.

4.2 评价准则

在AVEC2013和AVEC2014的抑郁症识别任务中,通过计算预测值和BDI-II之间的平均绝对误差(MAE)和均方根误差(RMSE)来评估提出方法的有效性.MAE可以表示为

$$E_{\mathrm{MA}}=\frac{1}{N}\sum_{i=1}^{N}\left|y_i-\tilde{y}_i\right|, \tag{4}$$

RMSE可以表示为

$$E_{\mathrm{RMS}}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\tilde{y}_i)^2}, \tag{5}$$

式中N表示数据样本数,y_i为数据库提供的BDI-II标签,$\tilde{y}_i$是第i个样本的预测值.

4.3 模型的训练参数

本文采用Python语言在TensorFlow平台上进行实验验证[19],模型的训练和测试是在拥有12 GB的NVIDIA Titan X进行,内存大小为64 GB.采用自适应矩估计(adaptive moment estimation,ADAM)优化器进行优化[20],其优点是允许根据模型权重更新的历史自动设置学习速率.学习率为0.000 001,迭代次数为100 000次.在初始输入上采用Dropout操作,在训练过程中随机选择的单元的激活设置为0,概率范围在0.5~1.0(每次迭代时随机设置).对于STA-ConvLSTM,65%的神经元被忽略,这样可以使方差降低,而且在一定程度上减少了过拟合.对于3D卷积层,由于滤波器通常很小(如5×5×5个滤波器的结果为125个神经元),如果以较高的百分比进行Dropout,则可能会对训练过程造成一定的特征减少,因此,只应用5%.但是,对于完全连接的层来说,由于隐藏单元的数量相对较高,应用50%的Dropout进行深层网络结构的训练.

对于深度网络的训练过程,权值初始化对最终结果的影响很大:如果网络中的权重开始太小,那么当信号通过每一层时会收缩,直到其太小而无法使用为止;如果权重开始过大,那么信号就会在穿过每一层时增长,直到其变得太大而无法使用为止.Xavier初始化确保权值在一个可以控制的范围之内,能够保证每一层输出的方差尽量相等,保持信号的数值在通过多层时在一个合理的范围之内.因此,使用Xavier初始化器进行权值的初始化.

对于STA-ConvLSTM来说,采用的激活函数是tan$h$和ReLU.由于tan$h$函数是有界的,二阶导数在无限趋近于0时,需要经过一段很长的时间,在空间注意力模型中得以采用,换句话说,就是保持时间不变性.ReLU函数能够快速收敛,因此,被采用在时序注意力模块中.抑郁症识别和分析是一个回归任务,采用MSE损失函数作为目标函数,公式为

$$J=\frac{1}{N}\sum_{i}^{N}(y_i-h_\theta(x_i))^2, \tag{6}$$

式中y_i是标签,$h_\theta(x_i)$是预测值,N是样本数.

4.4 抑郁症识别结果分析

为了更好地理解学习模型提取的特征,将在一个非常短的序列相对应的测试样本上的响应6张特征图进行可视化处理(见图6).由图6可知:第1个3D卷积后,面部图像的边界纹理特征已经充分地显示;第2个3D卷积以后,抽象出一些显著的对象(眼睛、鼻子等),这些对象是网络训练的对象,通过调整模型参数,能够自动识别出与抑郁症密切相关的模式.随着更深层次网络的训练和优化,可得到愈加抽象的与抑郁症密切相关的特征表示.

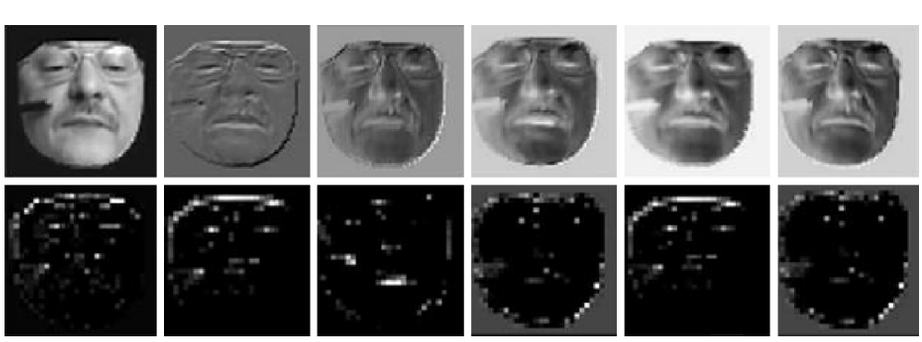


图6 一个短测试序列上的特征

基于AVEC2013和AVEC2014,提出的端到端抑郁症识别方法取得了优于大多数基于视频的识别结果.其优点在于充分有效地结合了3D-CNN、STA-ConvLSTM和SPP的优势.在该方法中,3D-CNN能够有效地刻画短时特征信息,STA-ConvLSTM通过注意力机制有效地表征了长时的空间和时间注意力高层信息,SPP能够将高层信息进行多尺度的变换,进而得到了紧致的最终特征表示,提高了抑郁症的识别率.

4.5　与现有方法的实验结果比较

本文提出了端到端的抑郁症识别方法,该方法能够对局部依赖信息和全局依赖信息进行有效的抽象和表征.特别是在STA-ConvLSTM模块中,采用了空间和时间的注意力机制进行关键信息的抽象和表征.将本文提出的方法与其他基于视觉特征的抑郁症识别方法进行比较.

在AVEC2013和AVEC2014测试集上的识别结果如表1所示.由表1可知,除了文献[4,6,26],本文的结果优于当前绝大部分基于面部视频的抑郁症识别方法的性能,与文献[12-13]相比,本文在AVEC2013和AVEC2014测试集上的RMSE误差分别减少了24.17%和5.43%,说明本文方法对基于视频的抑郁症识别的有效性.Zhu等[4]通过CASIA大型人脸数据库进行微调2个面部2D-CNN、光流2D-CNN,然后采用联合微调的方法进行抑郁程度的预测,取得了与目前基于视觉的抑郁症识别方法相当的识别结果;与Al等[6]的C3D-RNN方法进行比较,本文得到的RMSE和MAE较高,因为AI等先借助了预训练的C3D模型,然后在AVEC2013和AVEC2014数据库上进行微调,最后采用RNN进行时序信息的累积和建模,而本文直接在AVEC2013和AVEC2014数据库上并直接进行了模型的训练和评估,可能是样本量不足以及模型结构的复杂,从而导致RMSE和MAE没有降低.对于AVEC2014数据库,Dhall和Goecke[26]除了采用LBP-TOP、FV以外,在统计聚集方法中还使用了BoW进行二次聚集,BoW可以重新统计FV的一阶和二阶统计量中蕴含的模式信息,实现了抑郁症识别精度的提高.

综上所述,本文的主要贡献体现在:(1)设计了一个端到端的抑郁症识别方法,该方法包含了不同的深度学习技术,即3D-CNN、STA-ConvLSTM和SPP等;(2)提出的STA-ConvLSTM模块将局部短时的时空序列注意力信息进行全局的抽象和表征,有效地提高了抑郁症识别的准确率;(3)通过采用SPP对刻画不同尺度的信息,能够将显著的信息进行更加显著的表征,同时能够去除一定的冗余特征信息,最终得到更有价值的面部编码信息.

表1　不同数据库在不同方法的结果比较

方法	AVEC2013		AVEC2014	
	RMSE	MAE	RMSE	MAE
局部相位量化-三维正交平面[3]	10.27	8.22	—	—
深层卷积神经网络[4]	9.82	7.58	9.55	7.47
深层卷积神经网络[5]	8.19	6.30	8.39	6.21
三维深层卷积神经网络[6]	9.28	7.37	9.20	7.22
局部伽柏二值模式-三维正交平面[11]	11.19	9.14	—	—
局部相位量化[12]	13.61	10.88	—	—
局部伽柏二值模式-三维正交平面[13]	—	—	10.86	8.86
时空兴趣点[22]	10.45	—	—	—
局部伽柏二值模式-三维正交平面[23]	—	—	10.83	8.32
局部二值模式、边缘方向直方图、局部相位量化[24]	—	—	10.50	8.44
局部伽柏二值模式-三维正交平面、局部相位量化[25]	—	—	10.27	8.20
局部二值模式[26]	—	—	8.91	7.08
本文结果	10.32	8.60	10.27	8.60

注:—指无数据;RMSE指均方根误差;MAE指平均绝对误差.

5　结束语

本文提出了一种端到端的抑郁症识别方法,其采用3D-CNN提取局部时空信息,使用STA-ConvLSTM从局部时空信息中抽象长期依赖的注意力高层语义信息,并利用SPP进行特征尺度不变转换,得到了抑郁程度的估计.在AVEC2013和AVEC2014数据库上的实验结果表明,本文方法的性能得到了优于当前绝大部分基于面部视频的抑郁症识别方法.但是,该方法的主要局限性在于计算量大,需要学习的参数非常多,需要优化的参数量很大.在未来的工作中,应尝试减少网络参数,减轻网络负担,使之更加成为轻量化的端到端的抑郁症识别框架.

参 考 文 献

[1] STRATOU G,SCHERER S,GRATCH J,et al. Automatic nonverbal behavior indicators of depression and PTSD:exploring gender differences[C]//Affective Computing and Intelligent Interaction (ACII).2013 Humaine Association Confer-

ence on Geneva. Geneva:IEEE,2013:147-152.

[2] ALGHOWINEM S, GOECKE R, WAGNER M. Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors[J]. IEEE Transactions on Affective Computing, 2016, 9(4): 478-490.

[3] WEN L, LI X, GUO G, 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.

[4] ZHU Y, SHANG Y, SHAO Z, et al. Automated depression diagnosis based on deep networks to encode facial appearance and dynamics[J]. IEEE Transactions on Affective Computing, 2017, 9(4): 578-584.

[5] ZHOU X, JIN K, SHANG Y, et al. Visually interpretable representation learning for depression recognition from facial images[J]. IEEE Transactions on Affective Computing, 2020, 11(3): 542-552.

[6] AI 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.

[7] SONG S, JAISWAL S, SHEN L, et al. Spectral representation of behaviour primitives for depression analysis[J]. IEEE Transactions on Affective Computing, 2020, early access.

[8] AZHER UDDIN, JOOLEE J B, LEE Y K. Depression level prediction using deep spatiotemporal features and multilayer biLTSM[J]. IEEE Transactions on Affective Computing, 2020, early access.

[9] KAIMING H, XIANGYU Z, SHAOQING R, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis Machine Intelligence, 2014, 37(9): 1904-1916.

[10] SHI X, CHEN Z, HAO W, et al. Convolutional LSTM network: a machine learning approach for precipitation nowcasting [C]//SHI X. International Conference on Neural Information Processing Systems. Montreal: Curran Associates Inc, 2015: 802-810.

[11] MENG H, HUANG D, WANG H, et al. Depression recognition based on dynamic facial and vocal expression features using partial least square regression[C] //VALSTAR M. Proceedings of the 3rd ACM International Workshop on Audio/visual Emotion Challenge. Spain: ACM, 2013: 21-30.

[12] VALSTAR M, SCHULLER B, SMITH K, et al. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge[C] //VALSTAR M. Proceedings of the 3rd ACM International Workshop on Audio/visual Emotion Challenge. Spain: ACM, 2013: 3-10.

[13] VALSTAR M, SCHULLER B, SMITH K, et al. AVEC 2014: 3D dimensional affect and depression recognition challenge [C] //VALSTAR M. Proceedings of the 4th International Workshop on Audio/visual Emotion Challenge. Florida: ACM, 2014: 3-10.

[14] ZHAO G, PIETIKAINEN M. Dynamic texture recognition using local binary patterns with an application to facial expressions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(6): 915-928.

[15] CHAO L, TAO J, YANG M, et al. Multi task sequence learning for depression scale prediction from video[C]//JONATHAN G. 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). Xi'an: IEEE, 2015: 526-531.

[16] SZEDEGY C, IOEFF S, VANHOUCKE V, et al. Inceptionv4, Inception resnet and the impact of residual connections on learning[C]//SZEGEDY C. Thirty first AAAI Conference on Artificial Intelligence. California: ACM, 2017: 4278-4284.

[17] YANG L, JIANG D, SAHII H. Integrating deep and shallow models for multimodal depression analysis—hybrid architectures[J]. IEEE Transactions on Affective Computing, 2021, 12(1): 239-253.

[18] BALTRUSAITIS T, ROBINSON P, MORENCY L P. Openface: an open source facial behavior analysis toolkit[C]//Applications of Computer Vision (WACV). 2016 IEEE Winter Conference on Lake Placid. New York: IEEE, 2016: 1-10.

[19] ABADI M, BARHAM P, CHEN J, et al. Tensor flow: a system for large-scale machine learning[C]//ABADI M. Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation. Georgia: ACM, 2016: 265-283.

[20] KINGMA D P, BA J. Adam: a method for stochastic optimization[J]. arXiv: 1412.6980, 2014.

[21] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks[C]//GLOROT X. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Sardinia: PMLR, 2010: 249-256.

[22] CUMMINS N, JOSHI J, DHALL A, et al. Diagnosis of depression by behavioural signals: a multimodal approach[C]// VALSTAR M. Proceedings of the 3rd ACM International Workshop on Audio/visual Emotion Challenge. Barcelona: ACM,

2013:11-20.

[23] SIDOROV M, MINKER W. Emotion recognition and depression diagnosis by acoustic and visual features: a multimodal approach[C]//VALSTAR M. Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. Florida: ACM, 2014:81-86.

[24] JAN A, MENG H, GAUS Y F A, et al. Automatic depression scale prediction using facial expression dynamics and regression[C]//VALSTAR M. Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. Florida: ACM, 2014:73-80.

[25] KAYA H, ÇILLI F, SALAH A A. Ensemble CCA for continuous emotion prediction[C]//VALSTAR M. Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. Florida: ACM, 2014:19-26.

[26] DHALL A, GOECKE R. A temporally piece-wise fisher vector approach for depression analysis[C]//JONATHAN G. Affective Computing & intelligent Interfaces. Xi'an: IEEE, 2015:255-259.

Automatic depression estimation using 3D-CNN and STA-ConvLSTM from videos

HE Lang[1,2]

(1. School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an Shaanxi 710121;
2. Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an University of Posts and Telecommunications, Xi'an Shaanxi 710121)

Abstract: To represent the longterm temporal information from the facial image sequences, in this paper we propose to an end-to-end architecture, which adopts 3D Convolutional Neural Network (3D-CNN), Spatial Pyramid Pooling (SPP), and Spatial-Temporal Attention based ConvLSTM (STA-ConvLSTM) networks for depression recognition. On the test set of AVEC2013 and AVEC2014, this approach obtains the RMSE as 10.32 and 10.27, and the same MAE as 8.60, respectively, which are lower than those of most of the state of the art video-based depression recognition methods.

Keywords: depression; STA-ConvLSTM; 3D-CNN; ConvLSTM

(责任编辑:马田田)