基于增强特征和注意力机制的视频表情识别

李 飞[1],陈 瑞[2],童 莹[2],陈 乐[3]
(1. 南京工程学院 电力工程学院,江苏 南京 211167;
2. 南京工程学院 信息与通信工程学院,江苏 南京 211167;
3. 南京邮电大学 通信与信息工程学院,江苏 南京 210003)

摘 要:端到端的 CNN-LSTM 模型利用卷积神经网络(Convolutional Neural Network,CNN)提取图像的空间特征,利用长短期记忆网络 LSTM 提取视频帧间的时间特征,在视频表情识别中得到了广泛的应用。但在学习视频帧的分层表示时,CNN-LSTM 模型复杂度较高,且易发生过拟合。针对这些问题,提出一个高效、低复杂度的视频表情识别模型 ECNN-SA(Enhanced Convolutional Neural Network with Self-Attention)。首先,将视频分成若干视频段,采用带增强特征分支的卷积神经网络和全局平均池化层提取视频段中每帧图像的特征向量。其次,利用自注意力(Self-Attention)机制获得特征向量间的相关性,根据相关性构建权值向量,主要关注视频段中的表情变化关键帧,引导分类器给出更准确的分类结果。最终,该模型在 CK+和 AFEW 数据集上的实验结果表明,自注意力模块使得模型主要关注时间序列中表情变化的关键帧,相比于单层和多层的 LSTM 网络,ECNN-SA 模型能更有效地对视频序列的情感信息进行分类识别。

关键词:人脸表情识别;视频序列;自注意力机制;增强特征;卷积神经网络
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2022)11-0183-07
doi:10.3969/j.issn.1673-629X.2022.11.027

Video Facial Expression Recognition Based on ECNN-SA

LI Fei[1],CHEN Rui[2],TONG Ying[2],CHEN Le[3]
(1. School of Electric Power Engineering,Nanjing Institute of Technology,Nanjing 211167,China;
2. School of Information and Communication Engineering,Nanjing Institute of Technology,Nanjing 211167,China;
3. School of Telecommunications and Information Engineering,Nanjing University of Posts and Telecommunications,Nanjing 210003,China)

Abstract:The end-to-end CNN-LSTM model uses the convolutional neural network (CNN) to extract the spatial features of the image,and uses the long-term and short-term memory (LSTM) network to extract the temporal features between video frames. It has been widely used in video expression recognition. However,when learning the hierarchical representation of video frames,the CNN-LSTM model is complicated and prone to over fitting. Aiming at these problems,an efficient video expression recognition model with low complexity named ECNN-SA (Enhanced Convolutional Neural Network with Self-Attention) is proposed. Firstly,a video is divided into several video segments. The feature vector of each frame in one video segment is extracted by an enhanced CNN with global average pooling layer. Secondly,the self-attention mechanism is used to obtain the correlation between feature vectors,and the weight vector is constructed according to the correlation. The self-attention module with low computational complexity is used to focus on the frames of interest,which is greatly related to expression classification. The experimental results on CK+ and AFEW datasets show that the self-attention module makes the model mainly focus on the key frames of expression changes in the time series. Compared with the single-layer and multi-layer LSTM networks,the ECNN-SA model can classify and recognize the emotion information of the video sequence more effectively.

Key words:facial expression recognition;video sequence;self-attention mechanism;enhanced feature;convolutional neural network

0 引 言

人脸表情是人类传播内心情绪的重要方式,也是人类非语言的重要情感表达方式。人脸表情识别技术广泛应用于疲劳驾驶、在线教学、医疗等智能化人机交

收稿日期:2021-11-28 **修回日期**:2022-03-30
基金项目:国家自然科学基金青年项目(61703201,61701221);江苏省自然科学基金青年项目(BK20170765);江苏省未来网络科研基金项目(FNSRFP2021YB26);江苏省研究生科研创新计划(SJCX21_0945)
作者简介:李 飞(1995-),男,硕士研究生,研究方向为智能电网及信息技术。

互系统中,是模式识别和人工智能领域的研究热点。

基于视频序列的表情识别通过一个完整表情的运动过程能表达更丰富的表情变化信息,更具实际意义,也更具挑战。传统的视频序列表情识别方法有光流法[1]、隐马尔可夫模型[2]、运动历史图[3]和 LGBP-TOP(Local Gabor Binary Pattern Three Orthogonal Planes)[4]等。FAN 等[1]将梯度空间金字塔直方图扩展到时空域以获得三维特征,并将其与密集光流结合后得到时空描述符,用来提取人脸表情的空间和运动信息。局部二值算法(Local Binary Pattern,LBP)根据图像中每个像素与其局部邻域的点在亮度上的关系算出二值序,并对之编码后得到 LBP,最终图像的特征用多区域直方图来描述。这些方法多采用手工特征和浅层分类器,算法的鲁棒性较差。

随着深度学习技术在计算机视觉、图像与视频分析等领域的成功应用,卷积神经网络(Convolutional Neural Network,CNN)也被用于人脸表情识别,大大提高了识别精度[5-8]。Sun 等[5]在 EmotiW2015 竞赛中采用"Alex+RNN"的网络模型,其中 RNN 为循环神经网络(Recurrent Neural Network)的缩写,最终结果远远超过了竞赛的基准识别率。鉴于面部表情受到面部内不同区域姿势变化的影响,He 等[6]提出了一种多尺度特征提取器的 CNN,提高算法对于面部位置变化和尺度变化的鲁棒性。Jung 等人[7]采用一种模型视频序列中提取时变特征,另一种模型由单帧图像的面部关键点提取几何形状变化特征,从而联合微调网络的方法提高表情识别精度。多模态表情识别能进一步提高识别精度,如 Liu 等[8]在 EmotiW2018 竞赛中采用 DenseNet 网络处理音频,VGG-16 网络处理视频,长短期记忆(Long Short-Term Memory,LSTM)网络提取视频序列的运动特征,并用支持向量机(Support Vector Machine,SVM)提取关键点运动信息,将这些特征进行融合。这种多模态融合的表情识别方法虽然提高了识别精度,但方法复杂度高。

最近,CNN 结合 RNN 的网络架构用于视频序列表情识别,主要是利用了 RNN 的时间序列处理能力来获取视频序列时域动态信息,如 Chen 等[9]和 Khor 等[10]将 CNN 和 LSTM 网络级联起来,充分利用 CNN 强大的感知视觉表征与 LSTM 的时序处理能力;文献[11-12]采用 CNN 特征提取后,再用 RNN 完成特征的时序编码,结合时、空域信息完成基于视频的表情识别,提高了识别率。有研究表明,多层 LSTM 具有比单层 LSTM 更好的效果,如 Sutskever 等[13]提出的端到端序列学习方法中,级联了四层 LSTM,在长句子上表现良好,取得了良好的英法翻译性能;Irsoy 等人的研究[14]表明,与仅有单个隐藏层的 RNN 相比,具有紧凑结构的多层 RNN 计算效率更高。

除了上述研究成果,一部分研究者致力于将人类视觉系统的注意力机制(Attention Mechanism,AM)引入表情识别。梁斌等[15]结合多种 AM,提取更深层次的特征,在降低模型训练时间的同时,提高了目标表情的识别率。王晓华等[16]将 LSTM 网络堆叠起来,获得视频序列的分层表示,再用自注意力机制(Self-Attention Mechanism,SAM)描述层级的差异化,与单层 LSTM 相比,这种模型能更好地关注感兴趣层,获得更好的视频表情识别效果。文献[17]将时间特征和空间特征融合后,使用注意力机制进行特征加权,在 LSTM 网络中对加权后的特征进行训练和分类。这些方法能获得较好的识别结果,但深层 CNN 和多层 LSTM 网络的级联使得模型复杂度较高,且网络层数加深会出现梯度消失。因此,该文提出一种基于增强特征和自注意力机制的视频表情识别方法 ECNN-SA(Enhanced Convolutional Neural Network with Self-Attention),在 VGG-16 网络的中间层引出一条特征增强支路,并将其与骨干网络输出的深层特征相融合,用于获取不同层次的人脸表情特征,丰富表情信息。同时,用自注意力机制代替多层 LSTM 网络,不仅能有效学习序列内部的依赖关系,捕获内部结构和差异化的显著特征,而且自注意力机制主要是均值运算,避免了因网络层数加深而造成的梯度消失问题,大大加快了网络的训练速度。

1 视频表情识别网络模型 ECNN-SA

1.1 模型的总体框架

基于神经网络的视频表情识别过程包括视频获取、视频预处理、特征提取和分类识别。视频获取通常有两个来源:(1)实验室录制;(2)从已有的电影、电视节目中截取。该文的视频获取自网上的开源数据集:AFEW 和 CK+。其中,AFEW 是由不同电影中节选的视频片段,CK+是实验室中拍摄的受试者表情视频。视频预处理包括提取视频帧、人脸检测、人脸对齐、图像灰度转换,以及数据增强等。图像的特征提取和分类识别都可采用 CNN、RNN 等深度神经网络来完成。该文关注视频表情识别中的特征提取和分类识别,提出如图 1 所示的视频表情识别模型 ECNN-SA。ECNN-SA 模型由增强特征提取模块、自注意力机制(Self-Attention,SA)模块和回归模块构成。首先,视频序列输入 ECNN-SA 模型,由 ECNN 进行表情特征提取后送自注意力模块。具体实现时,一次处理连续的 10 个视频帧,经 ECNN 网络后,得到该 10 帧图像对应的特征向量 $\boldsymbol{x}_0,\boldsymbol{x}_1,\cdots,\boldsymbol{x}_9$。自注意力模块通过计算输入特征之间的相关性得到注意力权值 W_{sa},用这个

权重加权输入特征序列,从而获得差异化的显著性特征向量序列 $\boldsymbol{x}_0^*,\boldsymbol{x}_1^*,\cdots,\boldsymbol{x}_9^*$,这些向量的维度不变。接着,将 $\boldsymbol{x}_0^*,\boldsymbol{x}_1^*,\cdots,\boldsymbol{x}_9^*$ 序列的平均值输入回归模块。图 1 中的回归模块由 2 个全连接层(Fully Connection, FC)、1 个 ReLU 激活层、1 个 DropOut 层构成,其中 2 个 FC 层均完成特征映射的功能,第 1 个 FC 层的非线性由 ReLU 激活层来保障。为了防止过拟合,DropOut 层对特征值进行随机“灭活”。最终由 SoftMax 输出分类结果。

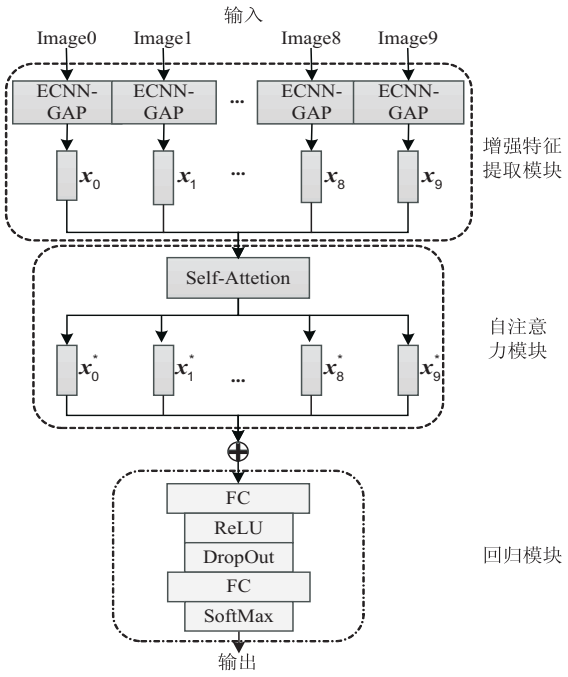

图 1　ECNN-SA 模型框架

1.2　增强特征提取模块 ECNN-GAP

为了降低模型的复杂度,考虑到全连接层有强大的拟合能力但模型复杂度高(占据了 VGG-16 大部分参数量),且容易过拟合,该文采用全局平均池化层(Global average Pooling,GAP)代替 FC 层完成对特征映射的降维。同时,为保证识别精度,引入增强特征支路,ECNN-GAP 模块如图 2 所示。

池化层方法若采用平均池化时,输出结果是滑动窗口中的数值求和取平均。池化层方法若采用 GAP 时,其窗口的尺度和特征映射的尺度相同,则无需全连接操作。由于全连接的参数太多,可用 GAP 代替。GAP 用特征图直接表示属于某个类的置信图。比如有 10 个类,最后输出 10 个特征图,每个特征图中的值加起来求平均值得到 10 个数字,这 10 个数字就是置信度。将这些平均值直接作为属于某个类别的置信度,再经过分类器进行分类。GAP 的使用可以大幅度减少模型的参数计算量。进一步,训练不同尺寸的图像时,由于 ECNN-GAP 模块输出的特征维度仅跟通道数有关,与尺寸无关,则不同大小的图像经 ECNN-GAP 后输出的特征维度都将保持一致。

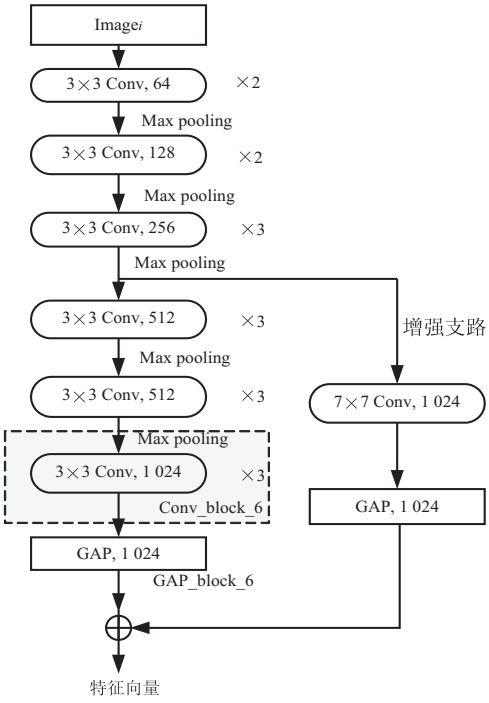

图 2　增强特征提取模块 ECNN-GAP

ECNN-GAP 模块中部分参数设置如表 1 所示,其中前五个卷积块与 VGG-16 相同。为增加模型深度获得更多的语义信息,增加了 Conv_block_6 卷积块。由表 1 中的参数设置可知,Conv_block_6 卷积块的通道数由 512 加大为 1 024,这样可使后续的 GAP_block_6 能提取出更丰富的特征向量。增强支路采用类似文献[18]的增强层,其中第一层 7×7 卷积,但去除了最大池化层和 1×1 卷积层,经 GAP 输出 1 024 维特征向量,最终 ECNN-GAP 模块输出的 2 048 维特征向量。

表 1　ECNN-GAP 模型部分参数

Layers Type	Output (Width×Height×Channel)	Params
Conv_block_6	7×7×1 024	3 层 3×3 卷积
GAP_block_6	1 024	-
Conv_enhanced	56×56×1 024	7×7 卷积
GAP_enhanced	1 024	1 024

1.3　自注意力机制模块

SA 使得人类视觉能够通过快速扫描全局图像找到感兴趣的目标区域,这个机制不仅能提高视觉信息处理的准确性,而且极大地提高了处理的效率。2017 年谷歌团队[19]提出的 SA 机制在机器翻译任务中获得了优秀的成绩。Fajtl J 等[20]将 SA 机制融入视频表情识别中,通过计算帧间相关性给每个视频帧打分,根据分数确定关键帧。如前所述,单层的 LSTM 在解决视频表情识别问题时,由于其仅传递一个层级的状态产生输出,从而对特征的表达能力显得不够。多层

LSTM 网络能提取不同级别的时间特征,比单层有更好的效果,但其时间复杂度较高。因此,该文采用 SA 机制代替多层 LSTM 网络,一方面通过 SA 模块学习序列内部的依赖关系,捕获内部结构,进而获取差异化的显著特征,另一方面随着网络层数的加深,由于 SA 模块采用的是均值运算,从而避免了梯度消失的问题,很大程度上提高了网络的训练速度。

SA 实质是一个将查询(Query)映射到正确的输入的过程,如图 3 所示,其中 Q 为查询, K 为键, V 为值,键 K 和值 V 之间有一个键值对(Key-Value pairs)表。查询 Q 、键 K 、值 V 和最终的输出都是向量,输出往往是一个加权求和的形式,权重由查询、键、值决定。源端中的元素由一系列的键值对构成,给定目标端中某个查询 Q ,每个键 K 对应到值 V 的注意力权重系数是通过计算查询 Q 和各个键 K 的相关性得到;再用 Softmax 函数对注意力权重进行归一化处理,将归一化注意力权重对 V 进行加权求和。自注意力机制可视为注意力机制的一种特殊情况,它不是应用在源端和目标端之间,而是源端内部元素之间或目标端内部元素之间发生的注意力机制,此时 $K = V = Q$,即:

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{1}$$

其中, $\boldsymbol{Q} \in R^{n\times d_k}$, $\boldsymbol{K} \in R^{m\times d_k}$, $\boldsymbol{V} \in R^{m\times d_v}$, d_k 为 $\boldsymbol{Q}$ 或 $\boldsymbol{K}$ 的维度, $\sqrt{d_k}$ 起调节作用,使得内积不至于太大。

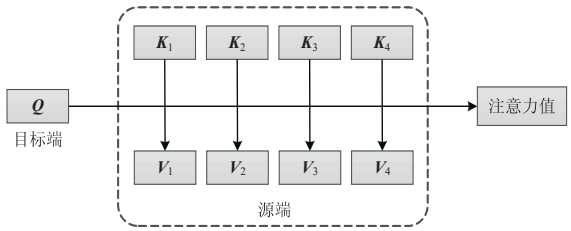

图 3 注意力机制

可见,SA 机制通过学习序列内容的依赖关系,进而捕获序列的内部结构,且计算简单。在视频表情识别中,结合 SA 机制来处理视频数据,让网络模型更加关注视频序列中差异性最大的帧,区分于视频的表情分类最相关的视频帧,更准确地识别面部表情。提出的 SA 模块如图 4 所示。

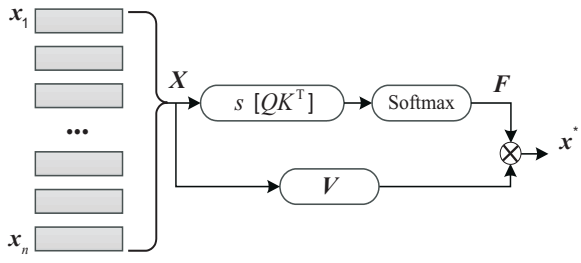

图 4 自注意力模块

图 4 中, $\boldsymbol{X} = [x_1, x_2, \cdots, x_n]$ 为 ECNN-GAP 网络输出的连续 n 帧图像的人脸表情特征向量, $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 的计算公式为:

$$\begin{cases} \boldsymbol{Q} = \boldsymbol{W}_q\boldsymbol{X} \\ \boldsymbol{K} = \boldsymbol{W}_k\boldsymbol{X} \\ \boldsymbol{V} = \boldsymbol{W}_v\boldsymbol{X} \end{cases} \tag{2}$$

其中, $\boldsymbol{W}_q$ 、$\boldsymbol{W}_k$ 和 $\boldsymbol{W}_v$ 为不同的网络权值矩阵。图 4 中,注意力权值矩阵 $\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}$ 描述了输入特征矩阵 $\boldsymbol{X}$ 中元素间相关性;超参数 s 可手工设置,用来抑制注意力权值大小。通过 Softmax 函数将注意力权重归一化到[0,1]区间,再与 V 相乘,得到差异化的显著特征矩阵 $\boldsymbol{X}^*$ 。这里, $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 都采用 2 048×2 048 的权值矩阵,相比全连接层,计算量大大降低。

2 实验结果与分析

为了验证 ECNN-SA 模型的有效性,在 CK+[21] 和 AFEW (Acted Facial Expression in the Wild)[22] 两个数据库上进行实验仿真。其中,AFEW 数据集为真实环境采集的非约束人脸表情数据库,样本受环境光照、姿态变化、遮挡、配饰、分辨率、拍摄角度、复杂背景等多种因素混合干扰,且因个体文化差异,受试者表现同类情感的程度也各不相同;CK+为实验环境采集的约束人脸表情数据库,样本中人脸正面姿态、无遮挡,且受试者根据实验要求夸张的表现各类情感。

该文提出的 SA 模块和回归模块的参数如表 2 所示。用来抑制注意力权重的超参数 s 设为 0.1。设一次同时处理 n 帧视频图像,则 SA 模块的输入为 n 个 2 048维特征向量,这些特征向量进行特征融合后输出 1 个2 048维特征向量,最后经过归一化处理、ReLU 函数激活和 DropOut 层之后,用一个 FC 层将前面提取的特征综合起来,对应到 7 类,最后 Softmax 输出人脸表情的分类结果。

表 2 Self-Attention 模块参数设置

网络层	输出大小	参数
Self-Attention	n ×2 048	K(2 048×2 048), V(2 048×2 048), Q(2 048×2 048), s(0.1)
Layer-norm	2 048	-
ReLU	2 048	-
DropOut	2 048	0.5
FC	7	-
Softmax	7	-

该模型训练时优化算法使用随机梯度下降算法,动量设置为 0.9。模型 VGG-16 卷积层部分加载在 SFEW 数据集和 FER2013 数据集上预训练的模型权重。模型初始学习率为 10^{-3},随着训练的过程衰减。输入图像统一预处理为 224×224 的灰度图像。实验代

码使用 Pytorch 编写在 Ubuntu 16.4 下完成,主机配备 2 块 NVIDIA GTX 1080Ti。

2.1　网络预训练和微调

由于 AFEW 数据库复杂程度高于 CK+数据库,因此,该文基于 AFEW 数据库进行网络预训练和微调。首先采用 VGG-FACE 权值作为骨干 CNN 网络的初始权值;然后用 SFEW 和 FER2013 中部分样本对自注意增强 CNN 网络进行微调;最后用 AFEW 的训练集及扩增的训练样本对自注意增强 CNN 网络进行训练,由此得到最佳网络参数。CK+数据库则在此网络上直接进行训练和测试。

2.2　AFEW 数据集上的结果及分析

AFEW 数据集[22]由不同电影中节选的视频片段组成,受试者具有自发的人脸表情,且受真实环境光照、姿态变化、遮挡、配饰、拍摄角度、分辨率、复杂背景等多种因素混合干扰,自 2013 年起作为 EmotiW 竞赛中的评估数据,每年组委会均会对 AFEW 数据库进行微调。

该文选择 2017 年竞赛数据 AFEW7.0 进行实验,将其分为三个部分:训练集(773 个样本),验证集(383 个样本)和测试集(653 个样本),其目的是为了确保三个数据集中受试者无重叠,由此验证人脸身份对人脸表情识别的影响。人脸表情标签有生气(anger)、厌恶(disgust)、害怕(fear)、开心(happiness)、中性(neutral)、悲伤(sadness)、惊讶(surprise)七种。AFEW 数据集中的连续表情图像如图 5 所示。

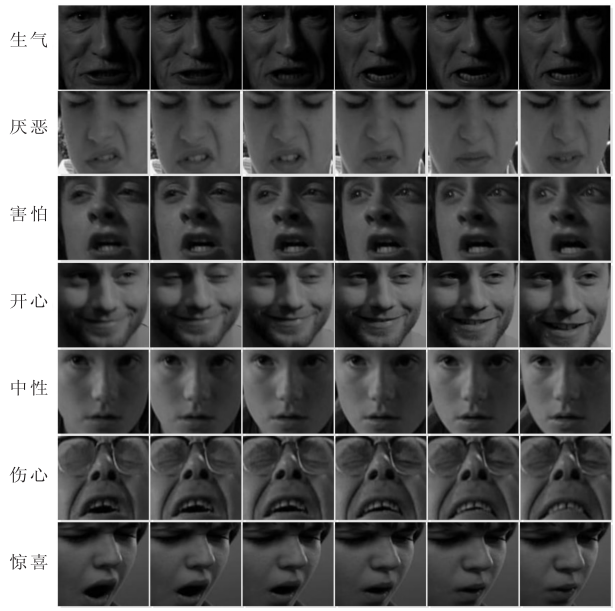

图 5　AFEW 数据集中的连续表情图像

调整 ECNN-SA 网络的结构和参数,在 AFEW 数据库上进行实验,得到的仿真结果如表 3 所示。

由表 3 可以看出,第 2 行 ECNN-LSTM 表示模型由增强 CNN 和 LSTM 网络构成,LBP 表示引入了传统的 LBP 特征。由第 2 行和第 3 行可以看出,引入传统的 LBP 特征可以提高准确率。第 4 ~ 11 行中,“CNN-SA(3 072, FC, s = 0.06)”的意思是,由 CNN 提取的 2 048 维特征向量与增强支路提取的 1 024 维特征向量一起构成输出的 3 072 维特征向量。在超参数 s 相同的情况下,输出向量的维度越低,识别效果越好,如第 4 行与第 7 行相比,第 5 行与第 6 行相比的结果表明,仅仅通过扩充通道数并不能提高识别效果。表 3 中,“FC”和“2×FC”分别代表 SA 模块后的 FC 层是一层还是两层。对比第 4 行和第 5 行,第 6 行和第 7 行的实验结果,可以看出一层 FC 的识别准确率更高,这是因为 FC 层的拟合能力太强,有可能导致过拟合,识别率不能提高反而会降低。

表 3　ECNN-SA 算法在 AFEW 数据集上的实验结果

	Model	Accuracy/%
1	Baseline	38.81
2	ECNN-LSTM (FC6, 7×7, LBP)	42.62
3	ECNN-LSTM (FC6, 7×7)	41.25
4	CNN-SA (3 072, FC,s=0.06)	40.64
5	CNN-SA (3 072, 2×FC,s=0.06)	39.76
6	CNN-SA (2 048, 2×FC,s=0.06)	40.91
7	CNN-SA (2 048, FC,s=0.06)	41.97
8	CNN-SA (2 048, FC,s=0.001)	41.14
9	CNN-SA (2 048, FC,s=0.1)	42.78
10	CNN-SA (2 048, FC,s=0.2)	42.25
11	CNN-SA (2 048, FC,s=0.5)	38.77

对于超参数 s 对识别准确率的影响,该文也进行了相应的实验。实验中,s 的取值在 0.001 到 0.5 之间选用多个数值进行实验。由表 3 中第 7 ~ 11 行的实验结果可以看出,相同条件下,在 s = 0.1 时获得最高的识别准确率 42.78%,比 Baseline 高出 3.97%。与 CNN-LSTM[18]相比,识别准确率提高了 1.53%;与带 LBP 特征的 CNN-LSTM 相比,准确率提高了 0.16%。综合以上的实验结果及分析,该算法不仅降低了计算复杂度,而且提高了识别准确率,算法可行且有效。

综上,可以得出结论:由于 FC 层具有强大的拟合能力,当增加其层数时,有时会导致模型过拟合,识别准确率下降。同时,通过单纯增加输出特征通道数提升网络识别性能,效果并不明显,当骨干输出维度 = 支路输出维度 = 1 024,超参数 s = 0.1 时,ECNN-SA 网络的性能最佳,识别准确率为 42.78%。

表 4 为采用 ECNN-SA 与传统 CNN-LSTM 网络端到端训练和测试一张样本的运行时间,表 5 为 AFEW 数据集上的混淆矩阵。由表 4 可以看出,当用 SA 机制代替传统 CNN-LSTM 中多层 LSTM 网络,且用 GAP 层代替 FC 层时,网络训练时间由原来的

40.34 ms 下降为 21.25 ms,下降了 47.32%,测试时间也下降 32.57%。同时,ECNN-SA 网络的识别准确率相比传统 CNN-LSTM 网络提高了 4.21%。

表 4 ECNN-SA 网络与 CNN-LSTM 网络的训练和测试时间对比 ms

CNN-LSTM 网络		ECNN-SA 网络	
训练时间	测试时间	训练时间	测试时间
40.34	8.26	21.25	5.57

表 5 ECNN-SA 模型混淆矩阵(AFEW 数据集) %

	生气	轻蔑	害怕	开心	中性	伤心	惊喜
生气	50	3.14	6.25	14.06	3.14	14.06	9.38
轻蔑	7.52	25	2.47	2.26	1.75	1.4	10.11
害怕	27.27	4.55	13.64	15.91	6.82	25	6.82
开心	4.84	1.61	0	79.03	4.84	6.45	3.23
中性	11.11	9.26	9.26	22.22	46.3	7.41	3.7
伤心	5	8.33	3.33	20	6.67	51.67	5
惊喜	28.89	4.44	2.22	15.56	17.78	15.56	15.56

由表 5 中的混淆矩阵可以看出,在 AFEW 数据集上,开心的表情识别率最好,其次是伤心和生气,其他表情的分类并没有取得非常好的效果,因为开心、伤心和生气相对于其他表情具有更加明显的特征,大多开心表情中的明显特征为嘴巴微张、嘴角翘起、眼睑收缩等,而害怕、轻蔑、中性等表情的特征有较为相似的特征,识别难度增大。另一方面,识别效果欠佳,原因主要有两个:

(1)该文采用的表情识别是单模态的,主要考虑连续面部表情序列相邻帧间的时间关系,如果采用多模态的方法,增加音频、文字等模式,能帮助提高表情识别率;

(2)AFEW 数据集是从不同电影中收集的视频剪辑,非常接近真实场景,包含各种头部姿势,演员的脸部遮挡,背景多变,是一个多模式数据库,因此在该数据集上识别结果较差。

2.3 CK+数据集上的结果及分析

CK+数据集[21]发布于 2010 年,是在 Cohn-Kanade 数据集基础上扩展来的,可以从网上免费获取,包含表情标签和 Action Units 的标签。这个数据集中包括 123 个受试者(subjects),593 个图像序列。每个图像序列的最后一帧都有 Action Units 的标签。图像序列中包含了从平静到表情表现峰值的图片。其中,来自 118 名受试者的 327 个图像序列被标记了七种基本情绪标签:生气(anger)、蔑视(contempt)、厌恶(disgust)、害怕(fear)、高兴(happiness)、悲伤(sadness)、惊讶(surprise)。CK+数据集中的连续表情图像如图 6 所示。

图 6 CK+数据集中的连续表情图像

由于 CK+没有给定训练集和测试集,将 327 个视频划分成长度为 10 帧的视频序列,共 978 个,取其中 80% 进行训练,20% 进行测试,交叉验证 5 次得到实验结果。

采用 2.2 节中在 AFEW 数据集上训练得到的具有最佳结构和参数的 ECNN-SA 网络(即骨干输出维度=支路输出维度=1 024,超参数 s=0.1)训练 CK+数据库,5 次交叉验证,得到的测试集实验结果如表 6 所示,对应的混淆矩阵如表 7 所示。由表 6 可以看出,提出的 ECNN-SA 网络达到最高识别率 97.95%,比其他网络 3DCNN-DAP[1]、STM-ExpLet[23]、DTAGN[7] 分别提高了 5.6%、4.07% 和 1.52%,比 CNN-LSTM[20] 网络提高了 2.03%。由此可见,在 AFEW 数据库训练得到的 ECNN-SA 网络在 CK+数据库上达到最佳识别性能。

表 6 CK+数据集上的测试结果

模型	准确率/%
3DCNN-DAP[1]	92.35
STM-ExpLet[23]	93.88
DTAGN[7]	96.43
端到端 CNN-LSTM[20]	95.92
特征增强模型[18]	97.47
ECNN-SA	97.95

表 7 CK+数据集上的混淆矩阵 %

	愤怒	轻蔑	厌恶	恐惧	快乐	悲伤	惊讶
愤怒	100	0	0	0	0	0	0
轻蔑	0	94.16	0	4.41	0	0	1.44
厌恶	0	0	98.3	0	1.7	0	0
恐惧	0	1.44	0	97.12	0	1.44	0
快乐	0	0	0	0	98.5	0	1.5
悲伤	0	0	0	0.68	0	99.32	0
惊讶	0	0	0.9	0	0	0	99.1

3 结束语

鉴于端到端 CNN-LSTM 网络用于视频表情识别时采用深层 CNN 提取空间信息和多层 LSTM 级联获取时间信息时,网络模型的复杂度较高且易发生过拟合,提出了一种高效、低复杂度的视频表情识别模型 ECNN-SA。使用改进后的 VGG-16 增强网络获取更多层次、更丰富的表情特征,用 SA 模块替代多层

LSTM 网络,对前后图像帧间的相关性进行加权求和,对表情最夸张的图像赋予较大的权值,最后由 Softmax 分类器输出表情类别。ECNN-SA 视频表情识别模型不仅更有效地获得了视频序列内部结构和差异化的显著特征,而且大大降低了网络的训练时间。

该方法是基于视频的面部表情识别,而人类的情感还包括语音、姿态动作和生理信号等,有效地融合音频、人脸关键点等其他模态的信息,对人脸表情进行多模态识别是下一步的研究工作。

参考文献:

[1] FAN X J, TJAHJADI T. A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences[J]. Pattern Recognition,2015,48(11):3407-3416.

[2] ELGARRAI Z, MESLOUHI O E, KARDOUCHI M, et al. Robust facial expression recognition system based on hidden Markov models[J]. International Journal of Multimedia Information Retrieval,2016,5(4):1-8.

[3] FAN X J,TJAHJADI T. A dynamic framework based on local Zernike moment and motion history image for facial expression recognition[J]. Pattern Recognition,2017,64:399-406.

[4] 郭振铎,徐庆伟,刘洲峰. 基于面部显著块动态信息的视频表情自动识别[J]. 计算机工程与设计,2017,38(6):1590-1594.

[5] SUN B,LI L D,ZHOU G Y,et al. Facial expression recognition in the wild based on multimodal texture features[J]. Journal of Electronic Imaging,2016,25(6):461-477.

[6] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]//Proc of the IEEE conference on computer vision and pattern recognition. Las Vegas:IEEE,2016:770-778.

[7] JUNG H,LEE S,YIM J,et al. Joint Fine-tuning in deep neural networks for facial expression recognition[C]//Proc of the IEEE international conference on computer vision. Santiago:IEEE,2015:2983-2991.

[8] LIU C,TANG T,LV K,et al. Multi-feature based emotion recognition for video clips[C]//Proc of the 20th ACM international conference on multimodal interaction. Boulder:ACM,2018:630-634.

[9] CHEN Rui,TONG Ying,LIANG Ruiyu. Real-time generic object tracking via recurrent regression network[J]. IEICE Transactions on Information and Systems,2020,E103-D(3):602-611.

[10] KHOR H Q,SEE J,PHAN R C W,et al. Enriched long-term recurrent convolutional network for facial micro-expression recognition[C]//Proc of the 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). Xi'an:IEEE,2018:667-674.

[11] SUN M C,HSU S H,YANG M C,et al. Context-aware cascade attention-based RNN for video emotion recognition[C]//Proc of the 1st Asian conference on affective computing and intelligent interaction. Los Alamitos:IEEE,2018:1-6.

[12] FAN Y,LU X J,LI D,et al. Video-based emotion recognition using CNN-RNN and C3D hybrid networks[C]//Proc of the 18th ACM international conference on multimodal interaction. New York:ACM,2016:445-450.

[13] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]//Proc of the 27th international conference on neural information processing systems. Cambridge:MIT Press,2014:3104-3112.

[14] LI M,IRSOY O,CARDIE C,et al. Physics-inspired neural networks for efficient device compact modeling[J]. IEEE Journal on Exploratory Solid-State Computational Devices and Circuits,2017,2017(2):44-49.

[15] 梁　斌,刘　全,徐　进,等. 基于多注意力卷积神经网络的特定目标情感分析[J]. 计算机研究与发展,2017,54(8):1724-1735.

[16] 王晓华,潘丽鹃,彭穆子,等. 基于层级注意力模型的视频序列表情识别[J]. 计算机辅助设计与图形学学报,2020,32(1):27-35.

[17] 何晓云,许江淳,史鹏坤,等. 基于注意力机制的视频人脸表情识别[J]. 信息技术,2020,2020(2):103-107.

[18] TONG Ying,CHEN Rui,LIANG Ruiyu. Unconstrained facial expression recognition based on feature enhanced CNN and cross-layer LSTM[J]. IEICE Transactions on Information and Systems,2020,E103-D(11):2403-2406.

[19] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]//Proc of advances in neural information processing systems. Long Beach:IEEE,2017:5998-6008.

[20] FAJTL J,SOKEH H S,ARGYRIOU V,et al. Summarizing videos with attention[C]//Proc of Asian conference on computer vision. [s. l.]:Springer,2018:39-54.

[21] LUCEY P, COHN J F, KANADE T, et al. The extended Cohn-Kanade dataset (CK+):a complete dataset for action unit and emotion-specified expression[C]//Proc of computer vision & pattern recognition workshop. San Francisco:IEEE,2010:94-101.

[22] DHALL A,GOECKE R,LUCEY S,et al. Collecting large, richly annotated facial-expression databases from movies[J]. IEEE Multimedia,2012,19(3):34-41.

[23] LIU M,SHAN S,WANG R,et al. Learning expressionlets via universal manifold model for dynamic facial expression recognition[J]. IEEE Transactions on Image Processing,2016,25(12):5920-5932.