·人工智能与模式识别· 文章编号：1000-3428(2020)06-0065-08 文献标志码：A 中图分类号：TP18

基于层次注意力机制的维度情感识别方法

汤宇豪，毛启容，高利剑

（江苏大学 计算机科学与通信工程学院，江苏 镇江 212013）

摘 要： 在连续维度情感识别任务中，每个模态内部凸显情感表达的部分并不相同，不同模态对于情感状态的影响程度也有差别。为此，通过学习各个模态特征并采用合理的融合方式，提出一种基于层次注意力机制的多模态维度情感识别模型。在音频模态中加入频率注意力机制学习频域上下文信息，利用多模态注意力机制将视频特征与音频特征进行融合，依据改进的损失函数对模态缺失问题进行优化，提高模型的鲁棒性以及情感识别的性能。在公开数据集上的实验结果表明，相比于卷积神经网络和长短时记忆网络等方法，该模型一致性相关系数指标明显提升，并且识别效率更高，可适用于大批量数据的维度情感识别。

关键词： 多模态；连续维度情感识别；注意力机制；特征融合；深度学习

开放科学（资源服务）标志码（OSID）：

中文引用格式： 汤宇豪，毛启容，高利剑. 基于层次注意力机制的维度情感识别方法[J]. 计算机工程，2020，46(6)：65-72.

英文引用格式： TANG Yuhao，MAO Qirong，GAO Lijian. Dimensional emotion recognition method based on hierarchical attention mechanism[J]. Computer Engineering，2020，46(6)：65-72.

Dimensional Emotion Recognition Method Based on Hierarchical Attention Mechanism

TANG Yuhao，MAO Qirong，GAO Lijian

（School of Computer Science and Communication Engineering，Jiangsu University，Zhenjiang，Jiangsu 212013，China）

【Abstract】 In continuous dimensional emotion recognition，the part of highlighting emotional expression varies in each modality，and different modalities also have different influence on emotional states. To address the problem，by learning modal features and fusing them in a reasonable way，this paper proposes a multimodal dimensional emotion recognition model based on Hierarchical Attention Mechanism（HAM）. Frequency attention mechanism is added to the audio modality to learn the context information in frequency domain，and the video features are fused with the audio features by using the multimodal attention mechanism. Then the problem of missing modalities is relieved by using the improved loss function to improve the robustness and emotion recognition performance. Experimental results on public datasets show that compared with methods such as Convolutional Neural Network（CNN）and Long Short Term Memory（LSTM）networks，this method has improved the Concordance Correlation Coefficient（CCC）index，and has higher recognition efficiency. It is applicable to dimensional emotion recognition of large volumes of data.

【Key words】 multimodality；continuous dimensional emotion recognition；attention mechanism；feature fusion；deep learning

DOI： 10.19678/j.issn.1000-3428.0054127

0 概述

情感是人类行为和思考的一种状态，随着人工智能技术的不断发展，人们更多地希望改变智能机器客观、冷静的特性，并进行深度开发使其具备人类一样的情感与思维，提供更人性化的服务。美国麻省理工学院 PICARD 教授根据情感在人类认知、决策、行动选择和语言学习等方面所起到的关键作用，于 1997 年提出了“情感计算”[1]的概念，其目的是通过赋予计算机识别、理解、表达和适应人的情感的能力来建立和谐人机环境，并使计算机具有更高更全面的智能。

随着情感计算需求的不断增加，需要识别的情

基金项目： 国家自然科学基金（61672267，61672268）。

作者简介： 汤宇豪（1994—），男，硕士研究生，主研方向为多模态情感识别；毛启容，教授；高利剑，硕士研究生。

收稿日期： 2019-03-07 **修回日期：** 2019-04-20 **E-mail：** 2211608031@stmail.ujs.edu.cn

感种类越来越多。传统的离散情感识别模型因为情感种类的局限性,在准确率和鲁棒性上达到了瓶颈。连续维度情感描述的是持续不断的情感状态,主要利用维度情感空间对情感状态进行建模和描述。这种方法采用空间中连续的数值来描述情感状态,每个情感状态对应多维空间中的一个点,每个维度对应情感的心理学属性,将描述情感阶段变化的离散情感转换为描述情感连续变化的维度情感。

本文提出一种层次注意力机制(Hierarchical Attention Mechanism,HAM)来学习音频模态中的频域信息和视频模态中的人脸位置信息,并将两者融合进行维度情感识别。该模型分为视频特征学习和层次注意力机制学习两个部分,通过频率注意力机制,计算音频不同频域对情感表达的贡献值并增强凸显情感流露部分特征的影响力,根据多模态注意力机制,分别计算两种模态对情感识别的贡献值并进行融合,以弥补单一模态信息表达不完整的缺陷。

1 相关工作

二维(arousal-valence)情感空间如图 1 所示,其中,横轴 valence 代表效价度,表示情感的积极与消极程度,纵轴 arousal 代表唤醒度,表示情感的激昂与低迷程度。通过设置效价度和唤醒度,可以表示出各种复杂细微的情感并加以区分,如欣喜若狂和怡然自乐描述了不同程度的愉悦之情,眉飞色舞和洋洋得意表达了两个褒贬不同的喜悦。二维情感空间因为其较简单的结构和丰富的情感表达能力,成为目前维度情感识别主要采用的维度空间。

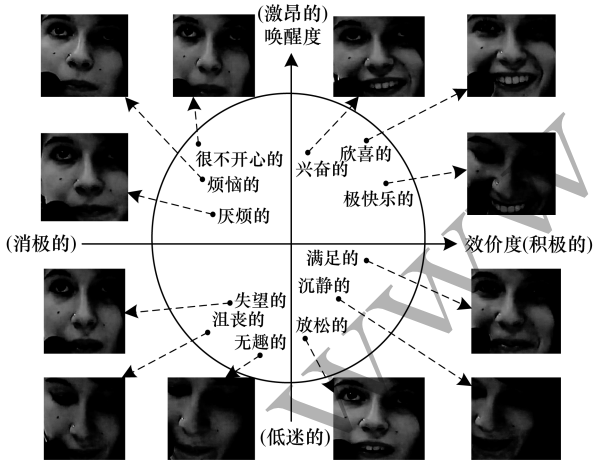


图 1 二维 arousal-valence 情感状态空间示意图

Fig. 1 Schematic diagram of two-dimensional arousal-valence emotional state space

早期的连续维度情感识别方法主要采用手工特征结合传统机器学习算法进行识别。文献[2]采用手工方法提取人脸表情特征,结合最大似然分类、似然空间估计等概率空间分类方法以及隐马尔科夫模型(Hidden Markov Models,HMM)实现维度情感识别。文献[3]采用支持向量机(SVM)算法和 k-近邻(KNN)算法对比维度情感识别效果。

随着深度学习的不断发展,卷积神经网络(CNN)和长短时记忆网络(LSTM)在维度情感识别领域得到应用。文献[4]使用手工方法和深度学习方法相结合的方式,首先将维度情感分为简单和复杂两个等级,使用隐马尔科夫模型对情感进行初步识别,然后在此基础上采用双向长短时记忆网络(BLSTM)学习时间上下文信息,识别效果优于传统机器学习方法。文献[5]采用时间池化的方式将多模态特征串在一起进行特征层融合并使用 LSTM 进行维度情感识别。文献[6]对音频和视频模态分别使用 BLSTM 进行识别,再运用线性支持向量回归(SVR)对识别结果进行决策层融合。文献[7]使用 3D 卷积神经网络学习特征上下文信息。虽然上述方法都取得了较好的效果,但是存在如下问题:

1)未考虑到人脸区域凸显情感表达的部分并不相同,如说话人微笑时,嘴部和眼部等凸显情感的部分较人脸边缘区域(头发、耳朵等)对情感识别影响更大[8]。此外,音频不同频域之间对情感识别的效果也有差异,同等处理高频和低频的特征并不合理,如激动时,高频域的特征相比于低频域的特征更能凸显此时的情感状态。

2)不同模态对于情感状态的影响程度是不同的,如说话人沮丧时,低沉的语调相比“面无表情”更能表征当前的情感状态。

3)已有模型所取得的高精确度主要源于数据库提供的手工特征以及在训练和测试模型时投入了高额计算成本。因此,如何采用更合理的方法进行多模态连续维度情感识别,成为当前的一个挑战。

近年来,注意力模型在自然语言处理、计算机视觉、语音识别等领域得到了广泛应用。文献[9]基于注意力模型构建了根据图像生成主题的模型。文献[10]提出基于 CRNN 与注意力机制相结合的语音识别模型。文献[11]提出将双向长短时记忆网络和注意力模型相结合的视频描述与语义分析的模型。其实注意力模型本质上是一种资源分配模型,主要目的是从众多信息中选择出对当前任务更关键的信息,提高模型的性能。以计算机视觉中的注意力模型为例,特征学习的瓶颈在于需要对整体图像处理[12],但是人类视觉只需要将视觉焦点集中在当前感兴趣的区域上,这一特点能够有效地减少人类视觉系统的带宽。因此,通过保留编码器(CNN、LSTM 等)对输入序列的中间输出结果,然后训练一个模型来对这些输入进行选择性的学习并且在模型输出时将输出序列与之进行关联。相比于采用多层

网络叠加或者决策层进行多模态融合的方式提高模型准确率,注意力模型使用更加简洁的结构学习对目标有利的特征,并将结果传递到下一层网络中进一步学习,简化了模型的复杂度,提升识别效果。

2　层次注意力机制维度情感识别方法

基于层次注意力机制(HAM)的多模态维度情感识别模型结构如图2所示,该模型主要分为视频模态特征学习和层次注意力机制两个阶段。在模型训练阶段,将训练视频输入到 HAM 模型中学习情感显著特征,将训练音频输入到频率注意力机制学习显著频域信息,然后利用多模态注意力机制将人脸特征和音频特征融合。在模型测试阶段,将测试视频输入到训练充分的 HAM 中,先提取人脸情感显著特征,再进行最终情感预测。本节首先对所提出的基于层次注意力机制维度情感识别模型进行概述,然后详细描述各个阶段的学习过程。

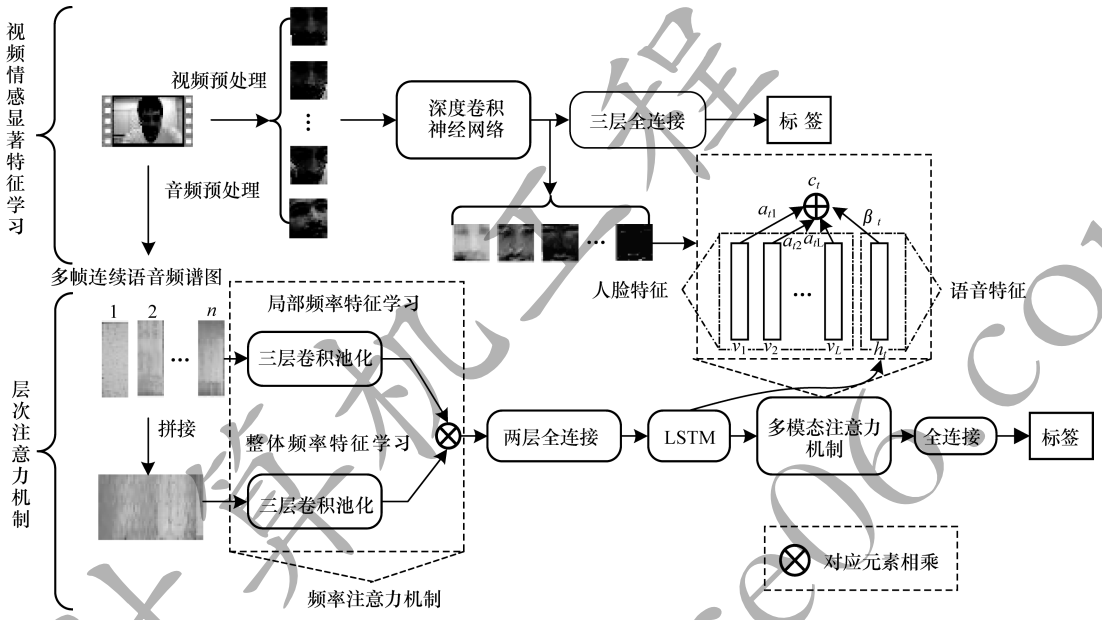


图 2　基于层次注意力机制的多模态维度情感识别模型结构

Fig. 2　Structure of multimodality dimensional emotion recognition model based on hierarchical attention mechanism

2.1　视频情感显著特征学习

首先通过视频预处理,将视频按照每 0.04 s 为一帧进行截取,采用数据库官方提供的人脸坐标对每帧图像中的人脸进行截取,再将所有人脸图片归一化到相同尺寸。然后采用深度卷积神经网络,包括 VGG、ResNet 和 Inception 对人脸图片序列进行特征学习。将深度卷积神经网络中的全连接层结构,改为三层全连接层,第 1 层包含 1 024 个节点,第 2 层包含 512 个节点,第 3 层包含 64 个节点。其中第 1 层、第 2 层全连接采用 relu 作为激活函数,第 3 层全连接采用 sigmoid 作为激活函数,主要目的是学习人脸特征中的高层特征,将特征中影响力较大的维度压缩到接近 1,影响力较小的维度压缩到 0.5,降低低层特征中的不稳定性。

相比传统方法中选取最后一层全连接层作为特征[13],本文采用最后一次卷积模块所得到的特征图作为人脸特征。全连接层是将池化后的特征拉直并进行压缩的过程,图像的位置信息和通道信息都被打乱,使用注意力模型学习到的特征贡献值只是形式化的参数,缺乏理论实际意义。而特征图则完整地保留了人脸的纹理和层次信息,如图 2 人脸特征所示,注意力模型更容易根据标签学习到人脸中的情感显著特征。特征图可以更直观地可视化出来,随着网络深度的增加,特征图会越来越稀疏,实验过程中相比于观察最后的识别准确率的变化,情感状态迁移给特征图带来的变化更明显。

2.2　层次注意力机制

本节首先通过音频预处理,按照每 0.04 s 为一帧提取音频频谱图。因为单帧语音频谱图展现的信息量较少,且缺乏上下文联系[14],所以以当前帧为基准,设置一个长度为 n 的滑动窗口,将前 $n-1$ 帧音频频谱图与当前帧频谱图进行拼接,作为当前帧的特征输入。然后滑动窗口以步长为 1 继续向后滑动采样。假设当前帧数少于 n 帧,比如第一帧,那么复制 $n-1$ 次第一帧进行补齐。由于前后帧与帧之间的变化较小,并且一般 n 取值小于 10,因此不会对实验产生波动性影响。

2.2.1　频率注意力机制

如图 2 所示,将提取好的单帧频谱图序列和整体频谱图分别输入到两个并行的三层卷积池化模块

中学习局部频率信息和整体频率信息。局部频率信息模块的输出激活函数为 sigmoid,把单帧频谱图特征值映射 0 ~ 1 之间,实际上这里的局部频率学习过程就是注意力矩阵的学习过程,主要学习的是不同频率之间特征的差异对整体频率信息带来的影响。将输出的局部频率特征进行拼接,与整体频率特征进行对应元素相乘,根据情感标签反向传播,来对整体频率特征中的各个维度进行选择性加强或者削弱。相比于只采用单帧音频频谱图作为输入,加入多帧频谱图可以使模型在特征学习阶段学习时间上下文信息,同时学习到帧与帧之间在频率上的差异,而不仅仅依赖于 LSTM 在后期进行时序构建。此外,传统的注意力模型往往需要在原有网络基础上增加一个分支来提取注意力权重,并进行单独训练,本文中的频率注意力机制在前向传播的过程中利用频率之间的差异性学习注意力矩阵,使得模型训练更加简单。

2.2.2 多模态注意力机制

将经过频率注意力机制处理过的音频特征经过全连接层输入到 LSTM 中学习时间上下文信息。假设 t 时刻音频特征为 x_t,前一时刻 LSTM 隐藏层输出为 h_{t-1},LSTM 门控函数为 $f()$,那么 t 时刻的隐藏层输出定义为:

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

假设 t 时刻人脸特征为 V,以 vgg19 为例,提取的特征是第 5 次卷积模块的输出,特征大小为 196 × 512,因此 V 的特征维数为 196,每个特征深度为 512,那么每维特征的注意力权重计算过程如下:

$$a_t = \text{softmax}(\boldsymbol{w}_h^{\mathrm{T}} \tanh(\boldsymbol{W}_v V + \boldsymbol{W}_g h_t)) \tag{2}$$

其中,$\boldsymbol{w}_h^{\mathrm{T}}$、$\boldsymbol{W}_v$、$\boldsymbol{W}_g$ 分别是注意力模型输入、视频特征和隐藏层输出的权重矩阵。经过注意力模型处理过的人脸特征为:

$$c'_t = \sum_{i=1}^{k} a_{ti} v_{ti} \tag{3}$$

将 LSTM 中的音频上下文信息和人脸特征融合,计算过程如下:

$$c_t = \lambda_t h_t + (1 - \lambda_t) c'_t \tag{4}$$

其中,λ_t 为 t 时刻音频特征的权重,$1 - \lambda_t$ 为 t 时刻视频特征的权重。

2.2.3 模态比例优化函数

在实验过程中发现,当说话人在说话时,镜头中并没有出现人脸,此时只有环境背景,如电脑仪器、桌子等,多模态注意力模型无法准确判断视频中是否存在人脸,只会根据特征中各维度大小依靠情感标签来反向传播分配相应的注意力权重,因此有可能在没出现人脸时依然给人脸特征分配了较大的贡献值。同理,镜头中出现了人脸,但是说话人并没有说话,而有可能是远处的录制视频人员发出的声音,但是仍然有可能给音频特征分配了较大的注意力权重,这样对情感识别造成了误导。在深度学习梯度下降过程中,会在原损失函数中增加 L2 正则化[15]函数来防止过拟合,其原理主要是通过增加辅助函数来限制原损失函数中无关参数的影响力,引导总损失函数反向求导的方向。因此,受 L2 正则化启发,本文采用增加辅助标签和辅助损失函数的方式在反向传播的过程中引导总损失函数梯度下降的方向,在出现极端情况(有人脸无声音,有声音无人脸)时,限制多模态融合比例的取值范围。针对音频模态,短时能量衡量了语音在某个时刻声音能量的强弱,由于远处录制人员和说话人距离较远,能量强度差异很大,因而通过设置能量阈值,低于阈值以下的能量强度设为 0,以此来规避掉远处录制人员声音的干扰。因此,提取音频短时能量并归一化到[0,1]作为音频辅助标签。针对视频模态,利用 opencv 中非常成熟的人脸检测库对人脸图片进行检测,检测到人脸则辅助标签置为 1,没检测到则置为 0。构造的辅助损失函数如下:

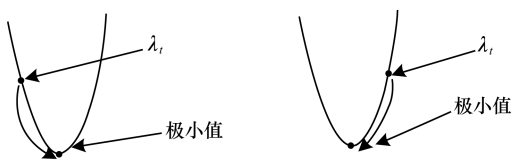
$$L_1 = \frac{1}{2}((m - \lambda_t)^2 + (n - (1 - \lambda_t))^2) \tag{5}$$

其中,m 为短时能量,是属于 0 ~ 1 之间的实数,n 为集合{0,1},表示是否检测到人脸。L_1 在 t 时刻对 λ 的梯度为:

$$\frac{\partial L_1}{\partial \lambda_t} = 2\lambda_t + n - m - 1 \tag{6}$$

当短时能量为 1 时,没有检测到人脸,即 $n = 0$,此时 $\frac{\partial L_1}{\partial \lambda_t} = 2\lambda_t - 2 < 0$,如图 3(a)所示,往极小值方向调整;当检测到人脸时,即 $n = 1$,但是短时能量 m 为 0,此时 $\frac{\partial L_1}{\partial \lambda_t} = 2\lambda_t > 0$,如图 3(b)所示,往极小值方向调整;当既有人脸又有声音时,即 $m = n = 1$;当既没有人脸也没有声音,即 $m = n = 0$ 时,$\frac{\partial L_1}{\partial \lambda_t} = 2\lambda_t - 1$ 接近于 0,对主损失函数影响很低。因此,总损失函数定义如下:

$$L = \frac{1}{2T}\sum_{n=1}^{T}(y'_t - y_t)^2 + L_1 \tag{7}$$



(a)有声音无人脸 (b)有人脸无声音

图 3 模态比例优化函数梯度示意图

Fig. 3 Schematic diagram of modal proportional optimization function gradient

3　实验结果与分析

3.1　数据库

为验证模型的识别效果,本文选用了 AVEC2016(International Audio/Visual Emotion Challenge and Workshop)挑战赛提供的数据库进行实验。AVEC2016 数据库是 RECOLA(Remote Collaboration and Affective Interaction)数据库的一个子集。数据库提供自然型的数据,是对参与视频会议的人进行录制得到的。数据库提供了训练集、验证集和测试集一共 27 个长度时间为 5 min 的视频,由6 个法国研究人员在 arousal 和 valence 两个情感维度上进行了标注,每隔 40 ms 进行一次标注,标注范围为 -1 ~1。每个视频长达 7 500 frame,最终每帧标签为 6 名研究人员标注结果取平均值。数据库官方强调了数据库构建的工作量并鼓励使用数据库的研究人员可以采用更合理的方法去提取特征。

3.2　实验设置

在视频特征学习阶段,本文采用 R 平方系数[16]作为特征学习的评估指标,其通过计算数据的变化来表征回归任务中预测值和标签值的拟合程度。R 平方系数越大,代表拟合程度越高,特征提取效果越好。R 平方系数函数如下:

$$R^2 = 1 - \frac{\sum (Y_{\_actual} - Y_{\_predict})^2}{\sum (Y_{\_actual} - Y_{\_mean})^2} \tag{8}$$

其中,$Y_{_actual}$是情感真实标签序列,$Y_{_predict}$是情感预测值序列,$Y_{_mean}$是情感真实标签序列的平均值。

在层次注意力机制训练与测试阶段,本文采用数据库官方提供的一致性相关系数(Concordance Correlation Coefficient,CCC)作为情感识别的评估指标,计算公式如下:

$$\mathrm{CCC}(x,y) = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{9}$$

其中,μ_x 和 μ_y 分别是情感预测值序列和情感真实标签序列的平均值,σ_x 和 σ_y 分别是情感预测值序列和情感真实标签序列的标准差,ρ 是 2 个序列之间的皮尔逊相关系数[17],计算公式如下:

$$\rho = \frac{\frac{1}{n}\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x\sigma_y} = \frac{C_{\mathrm{CCC}}}{\sigma_x\sigma_y} \tag{10}$$

在整个实验中,采用均方根误差(Root Mean Square Error,RMSE)作为损失函数,其定义如下:

$$\mathrm{RSME} = \sqrt{\frac{1}{n}\sum_i^n (x_i - y_i)^2} \tag{11}$$

其中,x_i 代表第 i 帧的情感预测值,y_i 代表第 i 帧的情感真实标签。

在视频特征学习阶段,选取 12 组视频作为训练集,6 组视频作为测试集。对于人脸特征学习,分别采用 Vgg19、ResNet34、ResNet50、InceptionV3 4 种经典深度 CNN 进行对比实验。由于以上 4 种深度 CNN 对图片大小要求基本都是 224 × 224 或者 299 ×299,因此批量训练数量设置为 150。在频率注意力机制模型中,使用三层卷积三层池化的卷积神经网络结构作为音频上下文特征学习模型,每帧音频频谱图大小为 24 × 120,输入窗口为 5 帧音频信息,因此输入频谱图大小为 120 × 120。第 1 层卷积核大小为 2 ×2,卷积核数量为 8。第 2 层卷积核大小为 3 ×3,卷积核数量为 16。第 3 层卷积核大小为 3 ×3,卷积核数量为 32。为了保证整体频率学习与频率局部学习特征图输出大小一致,卷积过程中采用全 0 填充,保持两者尺寸一致。池化尺寸全部设置尺寸为 2 ×2 的最大池化,步长为 1。

在验证 HAM 模型识别效果阶段,设置了 3 组对比实验:1)仅使用单帧音频频谱信息作为模型输入,在不使用频率注意力机制的前提下与视频特征在特征层融合,对比完整的 HAM 模型,比较 CCC 相关度系数;2)在使用频率注意力机制的情况下,与视频特征在特征层融合而没有在模态间使用注意力机制,对比完整的 HAM 模型,比较 CCC 相关度系数;3)在使用频率注意力机制和多模态注意力机制的情况下,对使用和未使用模态比例调整函数的实验效果进行对比。

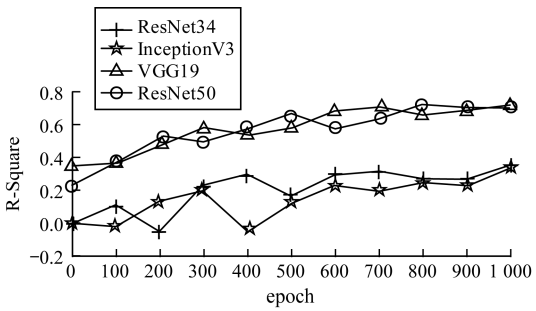
实验操作系统为 ubuntu18. 04,开发语言为 python3.6.2,深度学习框架为 tensorflow1. 8、keras 2.1 和 theano1.0.0,CPU 为英特尔至强 E5-2630V4 10 核 20 线程,内存为三星 ddr4 2400 16 GB ×8 (128 GHz),GPU 为英伟达特斯拉 P100 ×2 16 GB 显存,加速版本为 CUDA 9.0。在前期多次实验对比的情况下,为了保证训练充分,将 epoch 次数设置为1 000。梯度下降优化算法从 SGD、Adam 和 RMSProp 三者中选择。初始学习率设置为 0.000 5。为了更直观地对比训练和测试结果之间的差异,每训练一个 epoch 并在相应数据集上测试一次。

3.3　性能比较

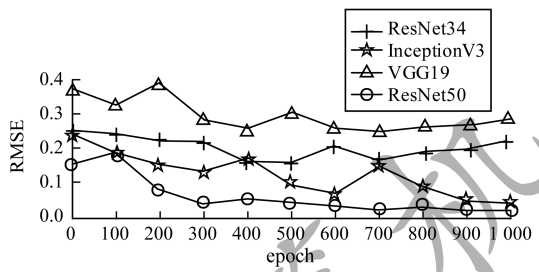
3.3.1　人脸情感显著特征学习效果对比

在人脸特征学习阶段,分别在 arousal 和 valence 2 个维度上采用 4 种深度卷积神经网络进行对比实验。特征学习结果分别如图 4 和图 5 所示,在 arousal 维度上,InceptionV3 和 ResNet50 在训练充分的情况下,R 平方系数都非常接近 0.73,两者损失几乎没有差异,但是 InceptionV3 相比 ResNet50 网络参数多达 24 734 048 个,单次 epoch 训练时间 54 s,而 ResNet50 单次 epoch 训练时间为 37 s,因此在 arousal 维度上采用 ResNet50 学习人脸特征。在

valence 维度上，ResNet34 的 R 平方系数达到了 0.62，损失也非常接近最低的 VGG19，而且网络结构相比于其他几种也更简单，单次 epoch 训练时间 31 s，因此在 valence 维度上采用 ResNet34 学习人脸特征。



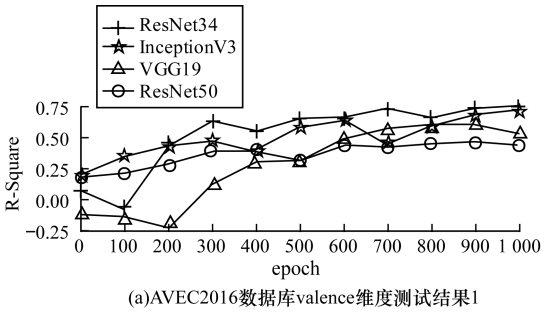
(a)AVEC2016数据库arousal维度测试结果1



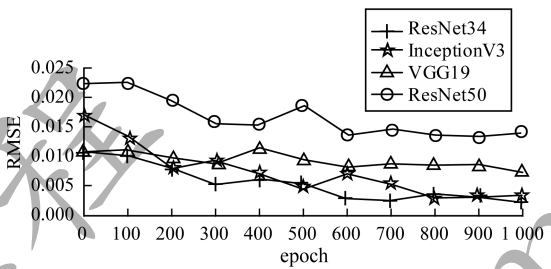
(b)AVEC2016数据库arousal维度测试结果2

图 4 arousal 维度视频情感显著特征学习结果

Fig. 4 arousal dimension video emotion salient feature learning results



(a)AVEC2016数据库valence维度测试结果1



(b)AVEC2016数据库valence维度测试结果2

图 5 valence 维度视频情感显著特征学习结果

Fig. 5 valence dimension video emotion salient feature learning results

3.3.2 层次注意力机制可视化

为更直观地展现层次注意力机制的识别效果，在测试模型阶段保存频率注意力机制和多模态注意力机制所计算出的注意力权重，叠加到原始人脸图片和频谱图上，生成热力成像图，并展示多模态注意力机制人脸特征权重分布图，如图 6 所示。

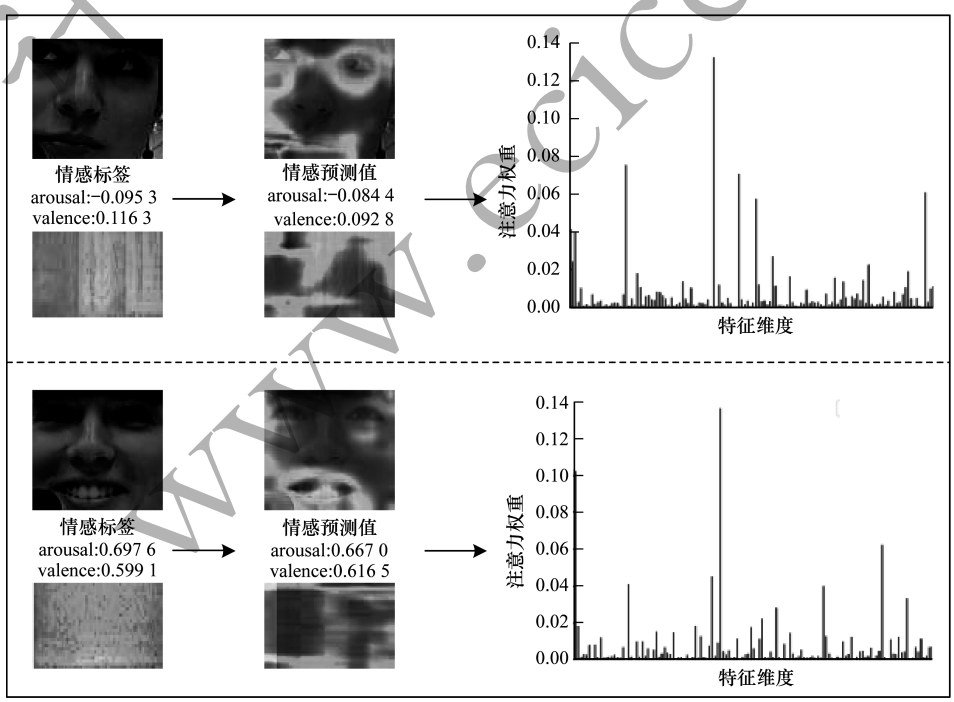


图 6 层次注意力机制可视化结果

Fig. 6 Visualization results of hierarchical attention mechanism

从图 6 可以看出，加入了层次注意力机制之后，人脸凸显情感表达的部位比如眼睛和嘴巴都被赋予了更大的权重，而边缘区域比如头发、耳朵等噪音的影响被削弱，语音信号中与当前情感流露更相关的

频率得到了加强,如图 6 所示说话人微笑时,高频域音频特征更加显著。通过注意力权重分布图可以看出,突出情感表达的特征往往集中在少数部分的几个特征中,这样有选择地加强这部分特征的影响力,减少了模型对其他无关特征或者影响力较小特征的关注,在面对维度更多的特征时,模型只需关注对当前识别贡献较多的特征。

3.3.3　层次注意力机制效果对比

在层次注意力机制学习阶段,分别对比了不使用层次注意力机制、仅使用层次注意力机制中的频率注意力机制而不使用多模态注意力机制、使用层次注意力机制和使用模态比例优化的层次注意力机制 4 种方法进行对比。实验中保存了 4 种方法在测试集上的最佳结果,并随机选取测试视频逐帧展开绘制预测曲线。在不使用层次注意力机制的情况下,预测曲线非常抖动,与标签曲线差异较大。使用了频率注意力机制的预测曲线,整体走向偏向标签曲线的发展趋势,相对稳定。使用了层次注意力机制的预测曲线,相比只采用频率注意力机制,预测曲线和标签拟合程度有了大幅提升,更加稳定。而加入了模态比例优化函数的预测曲线,在原有的结果上进一步优化,相对情感标签的拟合程度更高。

优化过的层次注意力机制在 arousal 和 valence 两个维度上的训练测试过程分别如图 7 和图 8 所示。在 arousal 维度上,当训练 epoch 达到 800 次时,此时训练集 CCC 已经超过 0.9,测试集也达到了 0.75左右,模型训练已充分,随着训练的继续,模型开始过拟合,测试效果下降。在 valence 维度上,训练和测试过程都较为抖动,训练 epoch 达到 700 次时,模型基本训练充分,但是测试结果并不稳定。无论是在 arousal 还是在 valence 维度上,RMSE 损失都有一定的波动。因此,在模型训练充分后,取测试阶段的 50 个 epoch 结果的平均值作为最终识别结果,最终在 arousal 维度上 CCC 为 0.732,在 valence 维度上 CCC 为 0.679。

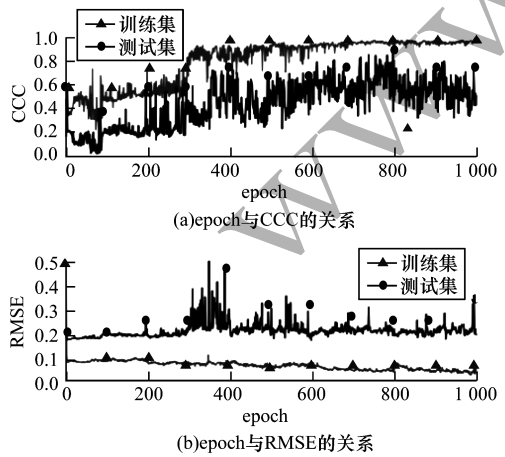


图 7　arousal 维度优化层次注意力机制训练测试过程

Fig. 7　arousal dimension optimization hierarchical attention mechanism training test process

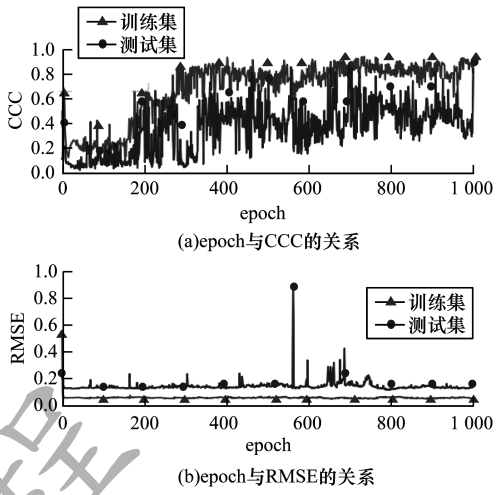


图 8　valence 维度优化层次注意力机制训练测试过程

Fig. 8　valence dimension optimization hierarchical attention mechanism training test process

具体对比实验结果如表 1 所示,相比于当前维度情感识别中的其他方法,层次注意力机制虽然在损失上逊色于最优结果,但是 CCC 相关度系数更能反映情感预测值和情感标签值的拟合程度。从表 1 可以看出,使用了频率注意力机制在 CCC 上已经超越了大多方法的结果,在此基础上构建的层次注意力机制的 CCC 表现最佳。最后对损失进行优化,CCC 在 2 个维度上分别达到了 0.732 和 0.679,说明经过优化的 HAM 模型可以更有效地提取音频和视频中的情感显著特征进行融合。

表 1　层次注意力机制与其他方法的维度情感识别结果对比

Table 1　Comparison of dimensional emotion recognition results between hierarchical attention mechanism and other methods

方法	唤醒度		效价度	
	RMSE	CCC	RMSE	CCC
文献[18]方法	0.263	0.431	0.174	0.478
文献[19]方法	0.188	0.535	0.121	0.463
文献[20]方法	0.168	0.533	0.114	0.354
文献[21]方法	—	0.617	—	0.467
文献[22]方法	—	0.615	—	0.530
文献[23]方法	0.152	0.683	0.337	0.642
Frequency Att 方法	0.229	0.653	0.216	0.601
HAM 方法	0.202	0.715	0.213	0.656
优化 HAM 方法	0.197	0.732	0.188	0.679

4　结束语

基于连续维度情感识别,本文提出基于层次注意力机制的维度情感识别方法。利用大量实验环境下的数据进行人脸情感显著特征和层次注意力机制两个部分的学习。实验结果表明,与目前的主流方法相比,本文方法使用注意力机制对所学习的特征利用上下文信息进行有选择的加强,简化了特征预

处理的过程,降低了情感无关因素的干扰,在连续视频与音频模态上的维度情感识别任务中取得了良好的效果。由于采用深度卷积神经网络学习人脸特征,且没有和层次注意力机制融合成一个模型,导致模型损失优化困难与特征学习不彻底,因此下一步将采用更合理的网络结构学习人脸特征,将特征学习和模型预测融为一体,并引入音频手工特征丰富音频信息与人脸特征融合,进一步提高模型识别准确率。

参考文献

[1] PICARD R W. Affective computing[M]. [S. l.]: MIT Press, 1997.

[2] NICOLAOU M A, GUNES H, PANTIC M. Audio-visual classification and fusion of spontaneous affective data inlikelihood space [C]//Proceedings of the 20th International Conference on Pattern Recognition. Washington D. C., USA: IEEE Press, 2010: 3695-3699.

[3] GRIMM M, KROSCHEL K. Emotion estimation in speech using a 3D emotion space concept[EB/OL]. [2019-02-20]. http://www.doc88.com/p-0877384901283.html.

[4] METALLINOU A, WOLLMER M, KATSAMANIS A, et al. Context-sensitive learning for enhanced audiovisual emotion classification [J]. IEEE Transactions on Affective Computing, 2012, 3(2): 184-198.

[5] CHAO Linlin, TAO Jianhua, YANG Minghao, et al. Multiscale temporal modeling for dimensional emotion recognition in video[C]//Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. New York, USA: ACM Press, 2014: 11-18.

[6] RINGEVAL F, EYBEN F, KROUPI E, et al. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data [J]. Pattern Recognition Letters, 2015, 66: 22-30.

[7] HUANG Jian, LI Ya, TAO Jianhua, et al. End-to-end continuous emotion recognition from video using 3D convlstm networks [C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2018: 6837-6841.

[8] ZHU Congxian. Research on deep learning-based speech emotion recognition method [D]. Nanjing: Southeast University, 2016. (in Chinese)
朱从贤. 基于深度学习的语音情感识别方法的研究[D]. 南京:东南大学, 2016.

[9] XU K, BA J, KIROS R, et al. Show, attend and tell: neural image caption generation with visual attention [EB/OL]. [2019-02-20]. https://arxiv.org/abs/1502.03044.

[10] CHOROWSKI J, BAHDANAU D, SERDYUK D, et al. Attention-based models for speech recognition [J]. Computer Science, 2015, 10(4): 429-439.

[11] GAO Lilian, GUO Zhao, ZHANG Hanwang, et al. Video captioning with attention-based LSTM and semantic consistency [J]. IEEE Transactions on Multimedia, 2017, 19(9): 2045-2055.

[12] JING Chenkai, SONG Tao, ZHUANG Lei, et al. A survey of face recognition based on deep convolutional neural network [J]. Computer Applications and Software, 2018, 35(1): 223-231. (in Chinese)
景晨凯, 宋涛, 庄雷, 等. 基于深度卷积神经网络的人脸识别技术综述[J]. 计算机应用与软件, 2018, 35(1): 223-231.

[13] ZHANG Jiakang, CHEN Qingkui. CUDA technology based Recognition algorithm of convolutional neural networks[J]. Computer Engineering, 2010, 36(15): 179-181. (in Chinese)
张佳康, 陈庆奎. 基于 CUDA 技术的卷积神经网络识别算法[J]. 计算机工程, 2010, 36(15): 179-181.

[14] HAN Wenjing, LI Haifeng, RUAN Huanbin, et al. Review on speech emotion recognition[J]. Journal of Software, 2014, 25(1): 37-50. (in Chinese)
韩文静, 李海峰, 阮华斌, 等. 语音情感识别研究进展综述[J]. 软件学报, 2014, 25(1): 37-50.

[15] FRIEDMAN J, HASTIE T, TIBSHIRANI R. Regularization paths for generalized linear models via coordinate descent[J]. Journal of Statistical Software, 2010, 33(1): 1-22.

[16] ISRAELI O. A Shapley-based decomposition of the R-square of a linear regression[J]. Journal of Economic Inequality, 2007, 5(2): 199-212.

[17] ADLER J, PARMRYD I. Quantifying colocalization by correlation: the Pearson correlation coefficient is superior to the Mander's overlap coefficient[J]. Cytometry Part A, 2010, 77(8): 733-742.

[18] VALSTAR M, GRATCH J. AVEC 2016: depression, mood, and emotion recognition workshop and challenge [EB/OL]. [2019-02-20]. https://www.researchgate.net/publication.

[19] CHAO Linlin, TAO Jianhua, YANG Minghao, et al. Long short term memory recurrent neural network based multimodal dimensional emotion recognition [C]// Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. New York, USA: ACM Press, 2015: 65-72.

[20] CHEN Shizhe, JIN Qin. Multi-modal dimensional emotion recognition using recurrent neural networks[C]//Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. New York, USA: ACM Press, 2015: 49-56.

[21] SCHUUER B, SCHULLER B. Multimodal sentiment analysis in the wild: Ethical considerations on data collection, annotation, and exploitation [EB/OL]. [2019-02-10]. http://www.1rec-conf.rogl.

[22] HUANG Z, STASAK B, DANG T, et al. Staircase regression in OA RVM, data selection and gender dependency in AVEC 2016 [C]//Proceedings of International Workshop on Audio/visual Emotion Challenge. New York, USA: ACM Press, 2016: 125-136.

[23] SUN Bo, CAO Siming, LI Liandong, et al. Exploring multimodal visual features for continuous affect recognition[C]//Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. New York, USA: ACM Press, 2016: 325-337.

编辑 索书志