计算机工程与应用
Computer Engineering and Applications
ISSN 1002-8331,CN 11-2127/TP

《计算机工程与应用》网络首发论文

题目：基于 ConvLSTM 网络的维度情感识别模型研究
作者：米珍美，赵恒斌，高攀
网络首发日期：2020-09-21
引用格式：米珍美，赵恒斌，高攀．基于 ConvLSTM 网络的维度情感识别模型研究．计算机工程与应用.
https://kns.cnki.net/kcms/detail/11.2127.tp.20200921.1509.017.html

基于 ConvLSTM 网络的维度情感识别模型研究

米珍美，赵恒斌，高　攀

石河子大学 信息科学与技术学院，新疆 石河子 832003

摘　要：学业情绪能够影响和调节学习者的注意、记忆、思维等认知活动，情绪自动识别是智慧学习环境中情感交互和教学决策的基础。目前情绪识别研究主要集中在离散情绪的识别，其在时间轴上是非连续的，无法精准刻画学生学业情绪演变过程,为解决这个问题,本文基于众包方法建立真实在线学习情境中的中学生学习维度情感数据集，设计基于连续维度情感预测的深度学习分析模型。实验中首先根据学生学习风格确定触发学生学业情绪的学习材料，并招募 32 位实验人员进行自主在线学习，实时采集被试面部图像，获取 157 个学生学业情绪视频；然后对每个视频进行情感 Arousal 和 Valence 二维化，建立包含 2178 张学生面部表情的维度数据库；最后建立基于 ConvLSTM 网络的维度情感模型，并在面向中学生的维度情感数据库上进行试验，得到一致性相关系数(Concordance Correlation Coefficient,CCC)均值为 0.581，同时在 Aff-Wild 公开数据集上进行试验，得到的一致相关系数均值为 0.222。实验表明，文中提出的基于维度情感模型在 Aff-Wild 公开数据集维度情绪识别中 CCC 相关度系数指标提升了 7.6%-43%。

关键词：连续维度情感识别；ConvLSTM；深度学习；学业情绪；维度情感数据库

文献标志码：A　　中图分类号：TP3-05

米珍美，赵恒斌，高攀. 基于 ConvLSTM 网络的维度情感识别模型研究.计算机工程与应用

MI Zhenmei, ZHAO Hengbin, GAO Pan. Research on the Dimensional emotion recognition model based on ConvLSTM Network. Computer Engineering and Applications

Research on the Dimensional emotion recognition model based on ConvLSTM Network

MI Zhenmei, ZHAO Hengbin, GAO Pan
College of Information Science & Technology, Shihezi University, Shihezi, Xinjiang 832003, China

Abstract: Academic emotions can affect and regulate learners' attention, memory, thinking and other cognitive activities. Automatic emotion recognition is the basis of emotion interaction and instructional decision in intelligent learning environment. At present, the research of emotion recognition mainly focuses on the recognition of discrete emotions, which is discontinuous in the timeline, and cannot accurately depict the evolution process of students' academic emotions. In order to solve this problem, this paper is based on the crowd-sourcing method to establish the dimensional emotional database of middle school students in the real online learning situation. And a deep learning

基金项目：国家自然科学基金项目（No.61965014）；兵团优秀青年教师人才专项项目（No.ZK20160201）。

作者简介：米珍美（1995-），女，硕士研究生，CCF 会员，主要研究领域为情感计算，深度学习，E-mail：mizhenmei@163.com；赵恒斌，男，硕士研究生，研究领域为计算机视觉；高攀（1981-），通讯作者，男，博士研究生，副教授，CCF 会员，研究领域为数据挖掘、图像处理和模式识别。

analysis model based on continuous dimensional affective prediction was designed. In the experiment, identify learning materials that Stimuli students' academic emotions according to students' learning styles firstly. And then 32 experimenter were recruited for independent online learning and collecting real-time facial images. Next, dimensional database obtained 157 students' academic emotion videos and 2178 students' facial expressions by the Two-Denationalization for each video emotion. Finally, a ConvLSTM net-based dimensional emotion model was established and tested on the dimensional emotion database for middle school students. The mean value of the concordance correlation coefficient (CCC) was 0.581. Meanwhile, the mean value of the uniform correlation coefficient was 0.222 after the experiment on Aff-Wild database. The experiment shows that the dimension-based emotion model proposed in this paper improves the CCC correlation coefficient index by 7.6% to 43% in the dimension-based emotion recognition of Aff-Wild database.

Key words: continuous dimension emotion recognition; ConvLSTM; deep learning; academic emotion; dimensional emotion database

1 引言

学业情绪不仅作用于学习者学习过程中产生的注意、记忆、决策等各个认知加工环节，而且影响学习者的学习动机和学习兴趣[1]。精准检测学习者学习状态是智慧学习环境的基础，也对实施个性化教育起着尤为重要的作用。学习者面部情感已成为教育情感计算中最常用的人工智能技术之一[2]，相比于离散情感模型在时间轴上是点式非连续的，维度情感模型是基于时间的一系列的数据，更能揭示数据的趋势性、规律性、异常性[3]。

目前基于维度情感计算研究主要针对人类的普通表情，而面向中学生学业情绪的研究却很少。分析维度情感预测研究，主要可分为回归和分类两类问题。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[4]。维度情感分类识别常用的算法有支持向量机(Support vector machine, SVM)[5]、隐马尔可夫模型(Hidden Markov model, HMM)[6]，维度情感预测常用的回归模型如支持向量回归(Support vector regression,SVR)等[7, 8]。随着深度学习的发展，循环神经网络（Recurrent Neural Network,RNN）以及其变体长短期记忆网络（Long Short-Term Memory，LSTM）也被用于维度情感检测[9, 10]。

Metallinou A[11]等结合隐马尔科夫模型与双向长短时记忆网络(BLSTM)进行维度情感识别，其结果优于传统的机器学习方法，余莉萍[12]通过改进 LSTM 在算法中引入注意力机制，将传统的遗忘门和输入门用注意力门进行替换，并在多个时刻的细胞状态 Fau Aibo 儿童情感数据语料库以及婴儿哭声情感数据库上得到比传统 LSTM 更为可观的识别结果。汤宇豪[13]则提出基于层次注意力机制的维度情感识别方法，将人脸信息与声音信息通过多层注意力进行有效融合，其结果表明模型在大规模的数据集中表现突出。Dimitrios Kollias[14]设计基于 CNN、CNN 和 RNN 相结合的模型进行情感维度识别，并在 CVPR 比赛中获得优异成绩。

虽然上述方法在维度情感方面取得了成功的应用，但是在面向中学生学业情绪识别上存在很大挑战：1.相比于基本情绪，学生在学习过程中产生的情感更加复杂，虽然研究者一直致力于识别更精准、更加丰富的人类情感，但其研究结果并不能直接应用于实际学习环境中；2.基于面部表情的维度情感识别更需要时空融合模型提取特征值，已有研究者提出 CNN 与 LSTM 堆叠相结合的方法,在时序模型 LSTM 阶段融合空间模型 CNN 进行时空特征

提取，忽略了 LSTM 时序建模中面部情感特征的学习。

因此，本文利用 ConvLSTM[15]网络进行维度情感识别，其不仅具有 CNN 刻画图像局部特征的能力，而且能够像 LSTM 一样建立时序模型，通过筛选有用的学生面部情感特征，解决 LSTM 网络无法处理冗余空间信息的问题。实验在自建的中学生学习维度情感数据库进行实现，并在 Aff-Wild 公开数据集[16]上进行试验，得到的相关系数均值为 0.222。实验表明，本文提出的基于维度情感模型在中学生学业情绪识别中 CCC 相关度系数指标提升了 7.6%-43%。

本文主要贡献有两点：（1）构建面向中学生的二维情感数据库；（2）通过经典深度卷积网络提取视频中的学生面部学业情绪，并将提取的特征输入 ConvLSTM 网络进行维度情感预测。

2 实验方法

2.1 维度情感预测

维度情感模型用几个取值连续的维度将情感刻画为一个多维信号，维度情感预测是对维度空间中每个维度的连续取值进行预测，通过对情感状态地实时标注来跟踪情感状态的演变过程.

基于 Arousal-Valence 二维情感空间从 Arousal、Valence 两个维度刻画情感，Valence 代表价效维度，表示情感的强烈和微弱程度。通过价效和唤醒两个维度可以区分更多细微的情感，每个人的情感状态可以根据价效维度和唤醒维度上的取值组合得到表征，这也使得机器能够更好地理解人的感情并做出精准的反应。

借鉴不同模态中基本情感维度的预测方法，宏观上模型分为面向中学生的不同学业情绪特征学习和维度情感预测两个阶段。在模型训练阶段，将待训练学生学业情绪视频输入到模型中学习情感显著特征。在模型测试阶段，将待测试维度情感预测视频输入到训练充分的算法模型中，先提取学生学业情绪的面部显著特征，再进行最终情感预测；实验中首先建立基于面向中学生的学业情绪数据集，通过 Arousal-Valence 二维情感空间描述学生的学业情绪，如图 1 所示部分学业情绪在维度空间中的表示；其次以中学生学业情绪数据集为基准筛选最优特征，并进行数据集与训练集的划分，其中训练集与测试集划分比例为 4:1；最后分析不同 CNN-LSTM 算法模型在情感维度中的预测结果，即使用 V、A 各维度的最优特征对算法模型进行训练，得到最好的模型，然后将测试集输入到训练好的算法模型中，得到待检测图像的 A、V 二维向量预测值。

图 1 二维(Arousal-Valence)情感状态空间

Fig.1 Two dimensional(Arousal-Valence) emotional state space

2.2 ConvLSTM 网络

LSTM 擅长时序数据的处理，但是如果时序数据是图像等三维图形，其有着丰富的空间信息并且每一个点与周围具有很强的相关性，普通的 LSTM 很难刻画这种空间特征,于是在 LSTM 的基础上加上卷积操作捕捉空间特征，对于图像的特征提取会更加有效。为了解决这个问题，Shi 等[11]设计了 ConvLSTM 网络，其将输入与各门之间的连接替换为卷积，从而融合 CNN 提取局部特征的能力和 LSTM 时序建模的能力于一起。传统 LSTM[17]有三个门：输入门，输出门，遗忘门，网络主要通过学习对这三者的控制来得到理想的结果，如果是多层结构的话，每个 LSTM 计算单元向上层传递的是 h 值。ConvLSTM 是 LSTM 的变体，主要是将 w 的权值计算变成了卷积运算，这样可以提取出图像的特征，如图 2 所示 LSTM 单元结构图。

LSTM 的输入、单元输出和状态都是一维向量，其关键公式如 1~5 所示，其中“$\circ$”表示 Hadamard 乘积：

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci} \circ c_{t-1} + b_i\right) \tag{1}$$

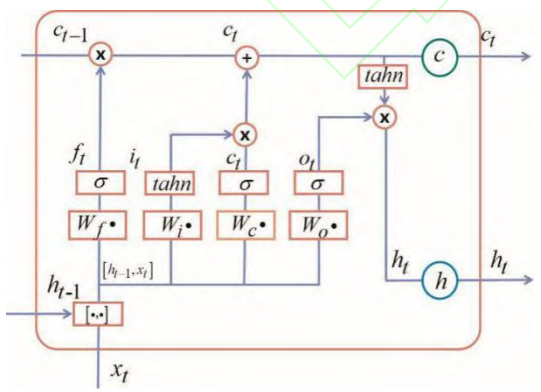


图 2 LSTM 单元结构图

Fig.2 LSTM cell structure

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf} \circ c_{t-1} + b_f\right) \tag{2}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \tag{3}$$

$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co} \circ c_{t-1} + b_0\right) \tag{4}$$

$$h_t = o_t \circ \tanh\left(c_t\right) \tag{5}$$

与传统网络不同，ConvLSTM 网络所有的输入 $X_{1,\cdots},X_t$ 细胞的输出 $C_{1,\cdots},C_t$，隐藏状态 $H_1,\ldots,H_t$ 以及输入门 i_t、遗忘门 f_t、输出门 o_t，均为三维向量，其最后两个维度代表行和列两个空间信息。在下面公式 6~10 中显示了 ConvLSTM 的关键等式，其中“$*$”表示卷积运算，“$\circ$”表示 Hadamard 乘积：

$$i_t = \sigma\left(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i\right) \tag{6}$$

$$f_t = \sigma\left(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f\right) \tag{7}$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh\left(W_{xc}X_t + W_{hc}H_{t-1} + b_c\right) \tag{8}$$

$$o_t = \sigma\left(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_{t-1} + b_0\right) \tag{9}$$

$$H_t = o_t \circ \tanh\left(C_t\right) \tag{10}$$

2.3 基于 ConvLSTM 网络的维度情感结构

基于 ConvLSTM 网络的维度情感模型结构如图 3 所示，通过卷积神经网络和 ConvLSTM 网络实现自动定位重要信息并对不同的帧分配不同的权重。首先,对中学生学业情绪视频进行预处理，为获取视频中学生面部特征，视频采样率 fps 值为 5，即每 0.2s 提取 1 帧，采用 Opencv 中的人脸特征模型对每一帧有效的学生学业情绪进行裁剪，并归一化到相同尺寸大小；接着将中学生学业情绪视频帧序列输入到由卷积神经网络构成的空间注意力网络中，随后将提取的卷积特征经过 ConvLSTM 解析后提取出长时间的序列特征，同时结合不同视频帧的时间信息生成视频的特征表示；最后，生成的特征表示经过全连接层和 tanh 激活函数，输出 V、A 二维向量预测值。

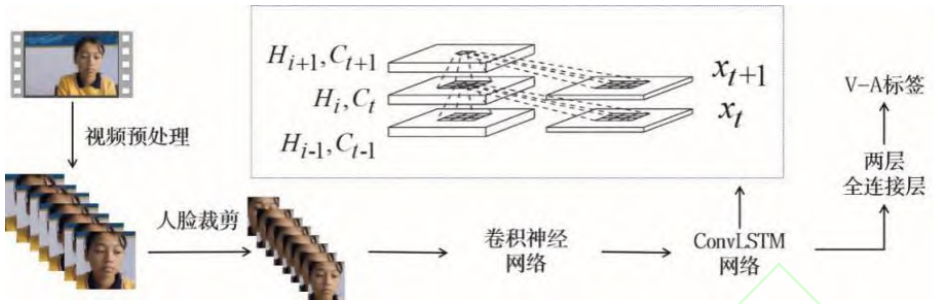


图 3　基于 ConvLSTM 网络的维度情感预测模型结构图

Fig.3　The structure of the dimensional affective prediction model of ConvLSTM network

实验中去除 VGG、ResNet 和 Inception 网络的全连接层，主要目的是学习中学生面部情绪中的高层特征，相比选取最后的全连接层作为特征，池化后提取的特征未经压缩和拉直，保留原始图像位置信息和通道信息，同时 ConvLSTM 网络要求保留面部学业情绪的特征矩阵。视频图像序列特征通过堆叠三层 ConvLSTM 网络，最终的预测结果由可能性最大的参数估算值决定，如公式 11 所示。通过多层叠加的 ConvLSTM 层，具有较强的时空表征能力，适用于维度情感等复杂问题的预测。

$$\tilde{x}_{t+1} = \arg\max_{x_{t+1}} p\left(x_{t+1} \middle| \hat{x}_{t-J+1}, \hat{x}_{t-J+2}, \cdots, \hat{x}_t\right) \quad (11)$$

3　数据库构建

实验通过模拟在线学习环境，实时采集学生学业情绪，创建了基于情感维度的中学生学业情绪数据库，数据库包括 157 个视频。实验采集了来自中学年龄在 12-18 的 32 名学生，其中男生 8 个女生 24 个 ,所有实验人员在实验开始之前均自愿签署了知情同意书。实验提前调查了被试学生所学知识以及知识水平，根据其学生学习特点，选择相应的知识内容，以使学生产生多样的学业情绪，图 4 显示了数据库中的一些帧，表征不同学生表现出不同学生表现出不同的学业情感。

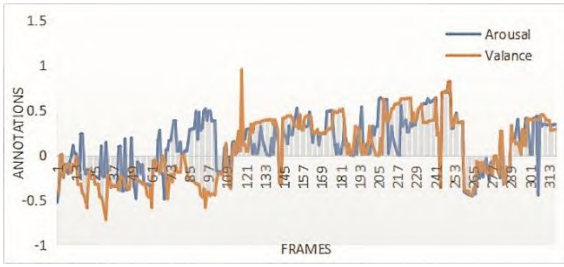


图 4　二维 Arousal-Valence 情感空间中的学生学业情绪

Fig.4　Two-dimensional Arousal-Valence academic emotion in the emotional space

实验结束，邀请 4 名标记人员依据二维 Arousal-Valence 情感空间和维度情感数据库[18]对情感视频进行标注。如图 5 和图 6 展示生成数据库中 Valence（a）和 Arousal（b）注释值的直方图。

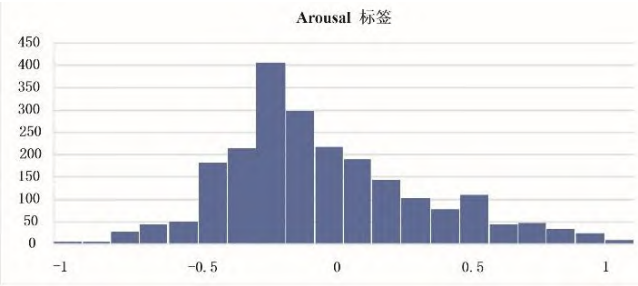


图 5　学业情绪数据库 Arousal 标签分布直方图

Fig.5　Academic emotional database Arousal label distribution histogram

3.1　数据预处理

为了更有效的提取面部特征 ,我们对视频进行预处理，通过 Peakutils 库提取视频帧，在每一帧中，我们使用 Adaboost 人脸检测算法进行人脸检测并进行裁剪[19]，再此过程中删除检测失败的帧，最终得到 2178 张学生面部表情帧。

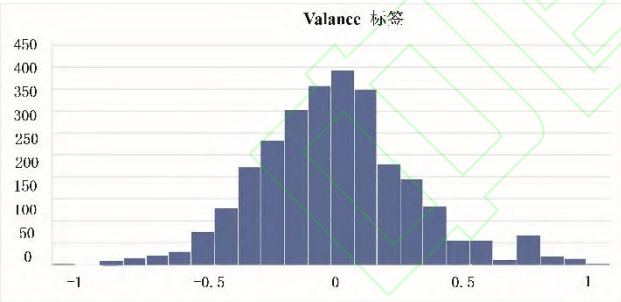


图 6　学业情绪数据库 Valence 标签分布直方图

Fig.6　Academic emotional database Valence label distribution histogram

3.2　数据标注

数据标注过程中使用了 ANNEMO[20]软件，一个基于情感和社会行为标注的 web 软件 ,其界面如图 7 所示。每个维度的标注过程如下：

（1）用户使用邮箱注册进行登录；

（2）同步所需标注的视频，用户可选择标注的视频；

（3）播放视频，通过左右移动标杆为视频标注 Arousal-Valence 值，其范围在[-1,1]之间，最后在数据库中存储每一帧生成相应的 Arousal-Valence 值。

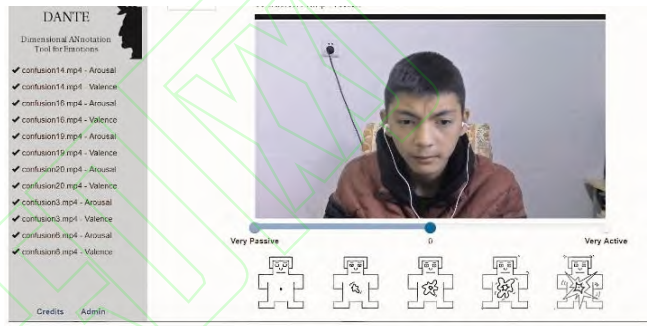


图 7　ANNEMO 标记界面图

Fig.7　The ANNEMO tag interface diagram

实验中选择四位标注人员进行视频情感维度标注，每位注释者均得到注释文档进行指导该任务的进行，该文档包括识别情绪 Arousal 和 Valence 的基础方法，标注人员通过对学业情绪状态的理解进行标注。其中在开始对每个学业视频标注之前，标注者观看了整个视频，以便对所显示的视频进行更为精准的标注。

3.3　注释统计分析

本文主要提供对标记者标记结果的详细分析，相比于离散情感模型，Arousal-Valence 情感模型可以用来更好地识别学生在学习过程中的情绪，通过定量和定性证明标记者标记结果的可靠性。图 8 中的散点图显示了 Arousal-Valence 值在在线学习中六种（专注，困惑，疲惫，厌烦、走神和愉快）常见的学业情绪的分布值。

从六种情感类别在 Arousal-Valence 空间中的分布情况可以看出：1)单一情感(如愉快)可产生多个 Arousal-Valence 值。这表明每一种类别的情感可能有不同的 Arousal-Valence 分布，这意味着传统离散情感类别可能不能准确地描述人的内心情感.2）情绪之间存在重叠，表明不同的情绪类别可能具有相似的 Arousal-Valence 分布。例如，某些“专注”和“愉快”图像的 Arousal、Valence 值非常接近。这表明每个人对语言特征都有不同的理解。在描述上，人类对情感的分类标记的一致性是相当差的。我们可以看到，从许多明确的词语中选择一种情感来描述一个人的情感是不容易的，因为有些情感标签之间有细微的差别，或者说情绪之间也有关系。

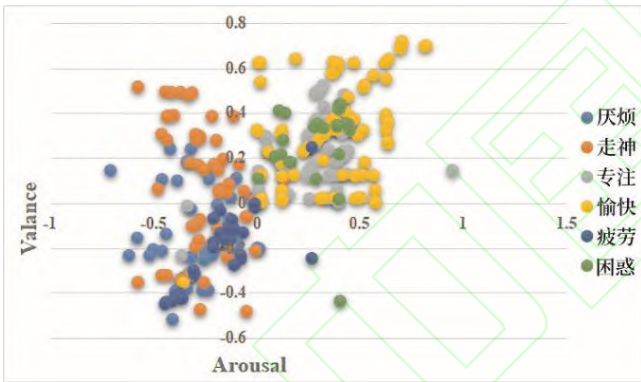


图 8　六种学业情绪在 Arousal-Valence 维度空间分布图

Fig.8　Six kinds of academic emotions in Arousal - Valence dimension space distribution

为进一步检验 Arousal-Valence 标签的质量，随机选取标记者的 500 个图像序列，我们使用了 Cronbach 的 alpha 方法评估数据的可靠性。在所有 Arousal-Valence 标签分数的 Cronbach’s alpha 值为 0.69，最小值值为 0.52。我们可以证明，不同标记者标注的标签之间的内部一致性是良好的。不同标记者之间的相关性分数 Pearson[21]相关系数为 0.46。

4　实验结果与分析

4.1　实验设置

实验中采用一致相关系数（Concordance Correlation Coefficient，CCC）和均方误差作为评估维度情感识别效能的评价指标。CCC 通过将两个时间序列(例如，所有标注视频和预测)的相关系数与它们的均方差进行缩放来评估它们之间的一致性。其取值范围为[-1，1]，其中+1 表示完全一致，而-1 表示完全不一致。CCC 的值越高，注释和预测之间的拟合越好。CCC 被定义为如公式 12 所示：

$$p_c = \frac{2 s_{xy}}{s_x^2 + s_x^2 + \left(\bar{x} - \bar{y}\right)^2} = \frac{2 s_x s_y p_{xy}}{s_x^2 + s_x^2 + \left(\bar{x} - \bar{y}\right)^2} \quad (11)$$

其中ρ_{xy}指皮尔逊相关系数（PCC）[21]，s_x和s_y分别为学生学习视频 Valance 或 Arousal 真实标签值和预测值，s_{xy}是相应的协方差值。

均方误差(Mean Square Error ,MSE)作为损失函数，其定义过程如公式 13 所示：

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - y_i\right)^2 \quad (12)$$

其中x和y分别是学生学习视频 Valance 或 Arousal 真实标签值与预测值，N是样本数量。MSE 的值越小，代表模型的预测能力越强。

实验中已标注的学生学习视频作为训练集和测试集，测试集与训练集数据之间相互隔离，其中训练集与测试集比例为 4：1，对比实验设置中分别采 Vgg19,ResNet34,ResNet50,InceptionV3 四种经典 CNN 进行提取学生面部学业情绪特征，并采用单向两层 LSTM 堆叠结构进行时序建模，tanh

作为激活函数，小批量梯度下降法，比较不同网络特征融合的预测结果，其中设置图片大小为224*224 或 229*229，后文称之为 CNN-LSTM 网络。在 ConvLSTM 网络中，使用三个 ConvLSTM 层进行特征学习，三层卷积核大小均为 7*7，卷积层中第一层卷积核数量为 32 ,第二层卷积核数量为 16，第三层卷积核数量为 8。为缩减模型计算量，在网络中使用最大池化层，设置尺寸为 4*4，图像矩阵边缘填充为“same”。

在验证 ConvLSTM 模型预测效果阶段，设置了三组对比实验：(1)使用 CNN-LSTM 网络进行维度预测，比较 CCC、MSE 相关度系数。(2)在使用 CNN-LSTM 情况下，分析 CNN-LSTM、CNN-GRU 的预测性能，比较 CCC、MSE 相关度系数。(3)将模型应用于 Aff-Wild 数据当中，分析其检测效果。

实现实验的操作系统为 Ubuntu16.04，深度学习框架为 Pytorch1.4，CPU 为 Intel 酷睿处理器，内存为三星 ddr4 2400 16G*2(32G)，GPU 为 GTX1080 Ti 显存，开发语言采用 Python3.5。在前期多次实验对比的情况下，为了保证训练充分，实验中比较了三种不同梯度下降优化算法 SGD、Adam 和 RMSProp，初始 epoch 次数设置为 1000，学习率设置为 0.0001。为了更直观的对比训练和测试的结果之间的差异，每训练一个 epoch 并在相应数据集上测试一次。

4.2 性能比较

4.2.1 不同 CNN-LSTM 网络预测结果对比

实验中依次使用四种深度 CNN 网络与 LSTM 进行结合分别对 Arousal 和 Valence 两个维度进行预测，结果对比如表 1 所示。实验中通过多次比较不同深度的 LSTM 网络，最终选择了两层的 LSTM 网络，并在网络最后连接 2 层全连接层实现 Valance 和 Arousal 两个情感维度的预测，实验中 VGG19 相对于 InceptionV3 和 ResNet50 网络参数都要少，但是其结果最佳，可能是因为数据量相对较少。

表 1　CNN-LSTM 网络预测结果对比

Table 1　Comparison of CNN-LSTM network prediction results

方法	Valance	Arousal	Mean Value
（A）CCC			
VGG16-LSTM	0.170	**0.219**	0.195
VGG19-LSTM	**0.361**	0.200	**0.281**
ResNet50-LSTM	0.172	0.163	0.169
InceptionV3-LSTM	0.195	0.166	0.181
（A）MSE			
VGG16-LSTM	**0.105**	0.130	0.117
VGG19-LSTM	0.119	**0.109**	**0.114**
ResNet50-LSTM	0.108	0.162	0.135
InceptionV3-LSTM	0.156	0.172	0.164

表 1 中实验针对学生维度情感数据库，结果显示，VGG 网络模型在 CCC 和 MSE 均表现最好，通过计算不同 CNN-LSTM 在 Arousal 和 Valence 的均值（Mean Value）可以看出 VGG19-LSTM 在 CCC 均值上至少高出其他 CNN-LSTM 模型 0.086；在 VGG 网络中 VGG19-LSTM 网络预测能力总体强于 VGG16-LSTM ,可以看出 VGG19-LSTM 网络对 Valance 维度的预测能力最佳，CCC 值高于 VGG16-LSTM 网络 0.191，CCC 均值高于 VGG16-LSTM 网络 0.086，并且 MSE 均值中低于 VGG16-LSTM 网络 0.003 ,因此适当增加网络深度可以增强网络对样本数据的学习能力，但并非越深的网络实验效果越好，由于训练样本数据量有限，

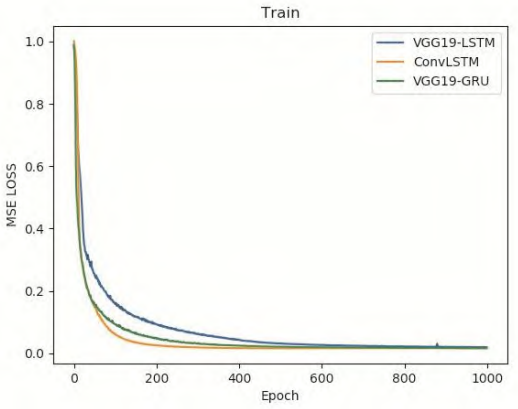


图 9　维度情感训练过程

Fig.9　Training process of dimensional emotional

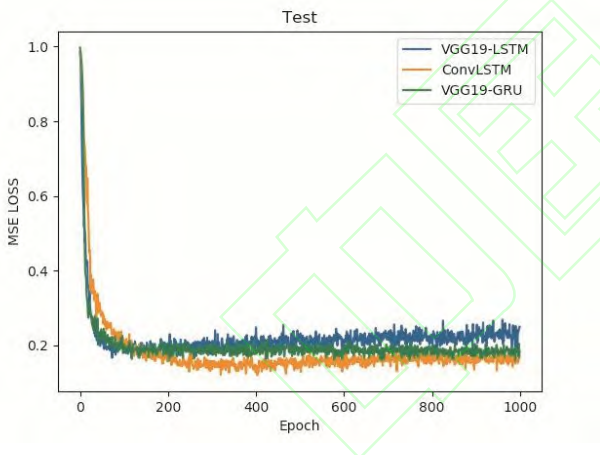


图 10　维度情感测试过程

Fig.10　Test process of dimensional emotional

ResNet50 网络在此数据集中模型没有取得好的效果。

4.2.2　ConvLSTM 上的性能分析

经典的 LSTM 网络中 state-state 采用全连接形式，而 ConLSTM 采用卷积的形式，分析 4.2.1 节结果，实验将使用 VGG19 网络提取特征并融合 ConLSTM 特征进行维度预测，分别比较了 VGG19-LSTM 与 ConLSTM 网络对学生学业情绪预测的能力，另外，实验比较了三种不同梯度下降优化算法 SGD、Adam 和 RMSProp。

ConLSTM 可以更好的学习图像输入的特征而不造成信息冗余。本文提出 VGG19-ConLSTM 结构不仅可以兼顾学生的面部表情特征，更能够克服时序数据对空间数据造成的冗余，又避免了 LSTM 无法实现对局部特征的刻画特点。

如图 9 和图 10 所示，三种网络 VGG19-LSTM、VGG-GRU 以及 ConvLSTM 在 Arousal 和 Valence 两
个维度上训练和测试时的 MSE 损失曲线，在最终训练模型的评估中 Arousal 和 Valence 的值均达到 0.9 以上，测试集中 ConvLSTM 表现最优，在 epoch 在 900 左右，模型接近于水平，图 10 中可以看出测试过程较为抖动，获取整个过程的 Arousal 和 Valence 两个维度的真实值和预测值，最终得到 Arousal 的 CCC 为 0.592，Valence 维度上 CCC 为 0.571。

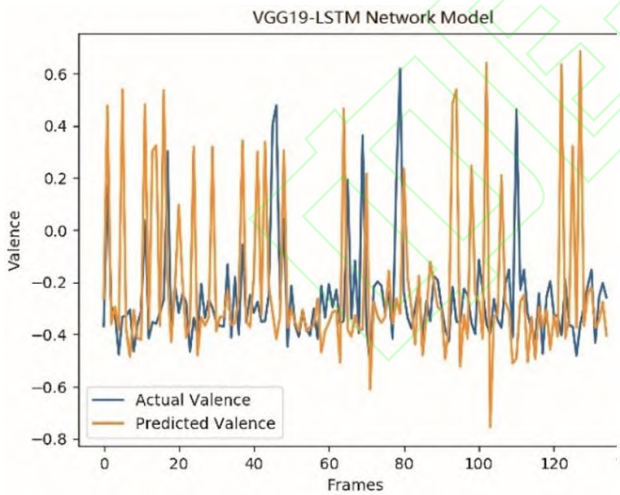
表 2　Aff-Wild 数据库实验结果比较

Table 2　Comparison of Aff-Wild database experiment results

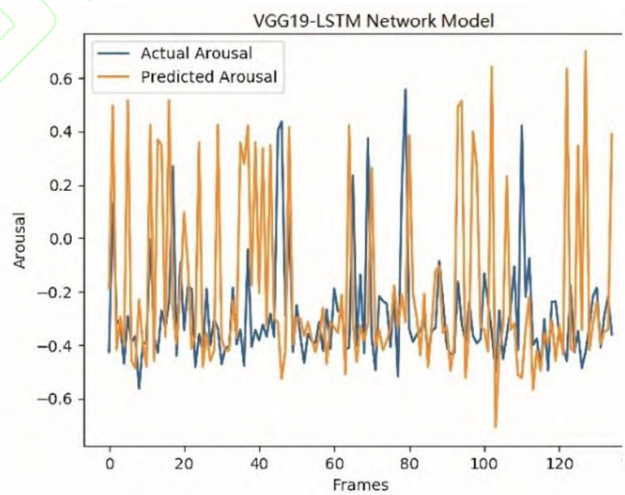
方法	Valance	Arousal	Mean Value
CCC			
MM-NET[22][22]	0.196	0.214	0.205
DRC-NET[23]	0.042	0.291	0.167
Baseline[24]	0.150	0.100	0.125
ConvlSTM	0.203	0.240	0.222
MSE			
MM-NET	0.134	0.088	0.111
DRC-NET	0.161	0.094	0.128
Baseline	0.130	0.140	0.135
ConvlSTM	0.051	0.097	0.074

如图 11 显示三种网络对测试集的预测能力，图中横坐标代表待测试帧，纵坐标为每一帧对应 Arousal 和 Valence 维度值，其中黄色线代表模型在 Arousal 和 Valence 两个维度上的预测值，蓝色线代表模型在 Arousal 和 Valence 两个维度上的真实值，图中可以看出 ConvLSTM 预测效果最好，VGG19-GRU 网络的预测效果相比于 VGG19-LSTM 网络较为逊色。因此，虽然 GRU 相对于 LSTM 模型结构复杂度低，需要更少的训练参数，但在数据集不同的情况下，模型预测能力是有所变化的。而 ConvLSTM 网络通过充分的提取空间特征并对特征进行筛选，充分提升预测网络能力。

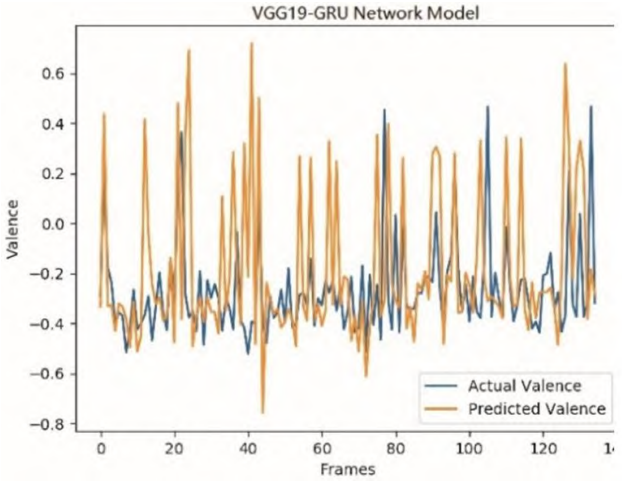
另外，本实验还将 ConvLSTM 模型应用在 Aff-Wild 数据库中进行测试，划分数据集为训练集和测试集，其中训练集和测试集比例为 4:1，具体对比实验结果如表 2 所示，相比于当前维度情感识别中的其他方法，ConvLSTM 虽然在损失上在损失远优于其他结果，但是 CCC 相关度系数更能反映情感预测值和情感标签值的拟合程度，可以看出，使用了 ConvLSTM 的网络在 CCC 均值上已经超越了大部分结果。CCC 在两个维度上分别达到了 0.203 和 0.240，这说明使用 ConvLSTM 网络在具有时空信息的维度情感预测中起到一定的效果。



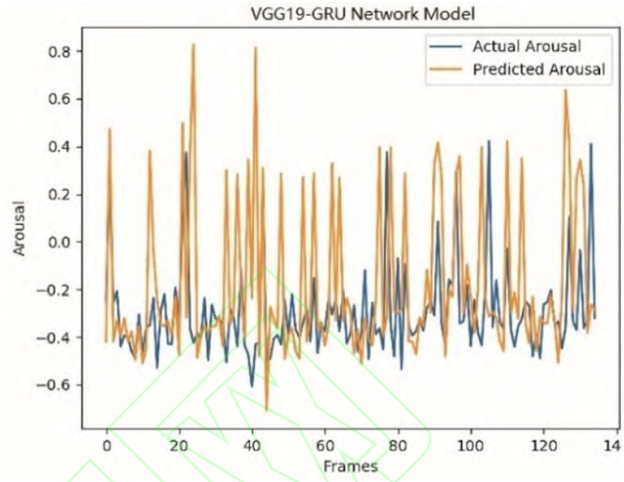
（a）VGG19-LSTM 模型对 Valence 的预测结果



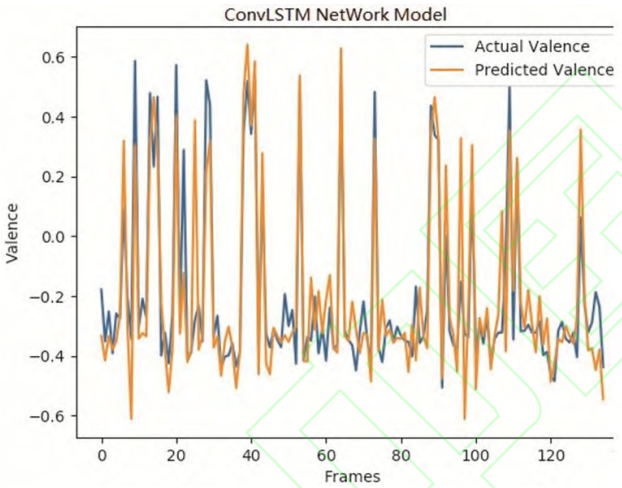
（b）VGG19-LSTM 模型对 Arousal 的预测结果



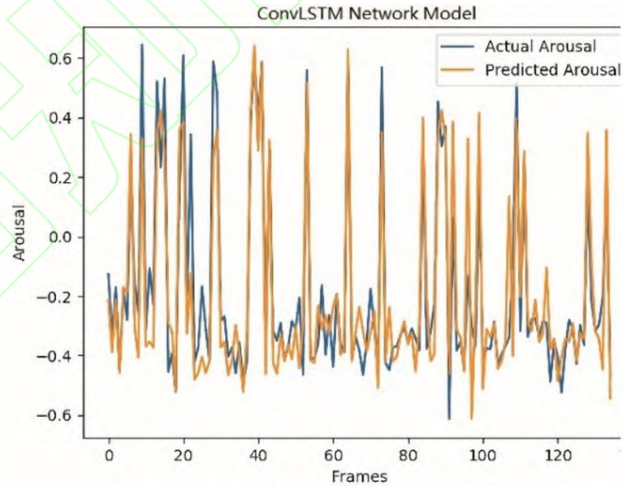
（c）VGG19-GRU 模型对 Valence 的预测结果



（d）VGG19-GRU 模型对 Valence 的预测结果



（e）ConvLSTM 模型对 Valence 的预测结果



（f）ConvLSTM 模型对 Valence 的预测结果

图 11　三种不同模型对 Arousal 和 Valence 的预测结果

Fig.11　Three different models predicted results for Arousal and Valence

5　讨论与结论

本文在 Vlence－Arousal 维度情感理论和教育心理学的基础上，实现了面向中学生的维度情感数据库，其中有 157 个学生学业情绪视频和 2178 张带有 Arousal 和 Valence 维度标签的学生面部表情。在此基础上，利用 ConvLSTM 网络能有效处理时空信息的能力设计维度情感预测模型，实现了面向学生学业情绪的维度情感预测。通过实验证明 ConvLSTM 与其他 CNN-LSTM 网络相比，在一致性相关系数和均方误差标准方面，均能提供最佳的

Vlence－Arousal 估计性能。实验结果表明，将 ConvLSTM 网络应用于面向中学生的维度情感预测是具有较好效果，为了测试模型预测能力，本文还在 Aff-Wild 公开数据集上进行试验，与目前的方法相比，本实验将学生面部的局部特征与其时间信息进行充分融合，减少数据冗余，识别中 CCC 相关度系数指标提升了 7.6%-43%。

学生学业情绪的精准测量是学生进行个性化学习的重要依据，本文将深度学习应用于教育中，实现学生的学业情绪预测，是教育与人工智能融合的有力尝试。当然，由于数据量不够，难免在精度上有一些欠缺。因此，未来的研究方向首先应该扩大面向学生的维度情感数据库并将其他特征融入到学生学业情绪中，比如学生学习的音频信息，学生学习的文本日志信息以及学生的学习行为数据，相信这些信息会进一步提高模型的预测能力。

参考文献

[1] 徐振国, 张冠文, 孟祥增, 等. 基于深度学习的学习者情感识别与应用[J]. 电化教育研究, 2019,40(02):87-94.

Xu Z G, Zhang G W, Meng Y Z, et al. Learners' Emotion Recognition and Its Application Based on Deep Learning[J]. E-education Research, 2019, 40(02): 87-94.

[2] Wu C, Huang Y, Hwang J. Review of affective computing in education/learning: Trends and challenges[J]. British Journal of Educational Technology, 2015.

[3] 曹晓明, 张永和, 潘萌, 等. 人工智能视域下的学习参与度识别方法研究——基于一项多模态数据融合的深度学习实验分析[J]. 远程教育杂志, 2019,37(01):32-44.

Cao X M, Zhang Y H, Pan M, et al. Research on Student Engagement Recognition Method from the Perspective of Artificial Intelligence: Analysis of Deep Learning Experiment based on a Multimodal Data Fusion[J]. JOURNAL OF DISTANCE EDUCATION, 2019,37(01):32-44.

[4] 李霞, 卢官明, 闫静杰, 等. 多模态维度情感预测综述[J]. 自动化学报, 2018,44(12):2142-2159.

Li X, Lu G M, Yan J J, et al. A Survey of Dimensional Emotion Prediction Based on Multimodal Cues[J]. ACTA AUTOMATICA SINICA, 2018, 44(12): 2142-2159.

[5] Ningthoujam S D, Hemachandran K. Content based Feature Combination Method for Face Image Retrieval using Neural Network and SVM Classifier for Face Recognition[J]. Indian Journal of Science & Technology, 2017,10(24):1-11.

[6] Swain M, Sahoo S, Routray A, et al. Study of feature combination using HMM and SVM for multilingual Odiya speech emotion recognition[J]. International Journal of Speech Technology, 2015,18(3):387-393.

[7] Liu B, Binaykia A, Chang P, et al. Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang[Z]. 2017: 12, e179763.

[8] Gunes H, Schuller B. Categorical and dimensional affect analysis in continuous input: Current trends and future directions[J]. Image & Vision Computing, 2013, 31(2): 120-136.

[9] Pei E, Yang L, Jiang D, et al. Multimodal dimensional affect recognition using deep bidirectional long short-term memory recurrent neural networks[C]. affective computing and intelligent interaction, 2015: 208-214.

[10] Zhang L, Zhang J. Synchronous prediction of arousal and

valence using LSTM network for affective video content analysis[C]. international conference on natural computation, 2017: 727-732.

[11] Metallinou A, Wollmer M, Katsamanis A, et al. Context-Sensitive Learning for Enhanced Audiovisual Emotion Classification[Z]. 2012: 3, 184-198.

[12] 余莉萍，梁镇麟，梁瑞宇. 面向儿童情感识别的改进LSTM[J]. 计算机工程, 2020:1-14.

Yu L P, Liang Z L, Liang R Y. Improved LSTM for Children's Emotion Recognition[J]. Computer Engineering, 2020, 1-14.

[13] 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, 2019:1-8.

Tang Y H , Mao Q R, Gao L J. Dimensional Emotional Recognition Based on Hierarchical Attention Mechanism[J]. Computer Engineering, 2019, 1-8.

[14] Kollias D, Tzirakis P, Nicolaou M A, et al. Deep Affect Prediction in-the-Wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond[J]. International Journal of Computer Vision, 2019,127(6):907-929.

[15] Shi X, Chen Z, Wang H, et al. Convolutional LSTM Network: a machine learning approach for precipitation nowcasting[C]. neural information processing systems, 2015: 802-810.

[16] Zafeiriou S, Kollias D, Nicolaou M A, et al. Aff-Wild: Valence and Arousal 'In-the-Wild' Challenge[C]. computer vision and pattern recognition, 2017: 1980-1987.

[17] Graves A. Generating Sequences With Recurrent Neural Networks[J]. CoRR, 2013,abs/1308.0850.

[18] Qinglan Wei, Sun Bo, He Jun, et al. BNU-LSVED 2.0: Spontaneous multimodal student affect database with multi-dimensional labels[J]. Signal Processing:Image Communication, 2017, 59168-181.

[19] Pham T M, Doan D C, Hitzer E. Feature Extraction Using Conformal Geometric Algebra for AdaBoost Algorithm Based In-plane Rotated Face Detection[J]. Advances in Applied Clifford Algebras, 2019,29(4):1-19.

[20] Ringeval F , Sonderegger A , Sauer J , et al. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions[C]//Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on. IEEE, 2013.

[21] Adler J, Parmryd I. Quantifying colocalization by correlation: The Pearson correlation coefficient is superior to the Mander\"s overlap coefficient[J]. Cytometry Part A, 2010,77A(8):733-742.

[22] Li J , Chen Y , Xiao S , et al. Estimation of Affective Level in the Wild with Multiple Memory Networks[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2017.

[23] Hasani B , Mahoor M H . Facial Affect Estimation in the Wild Using Deep Residual and Convolutional Networks[J]. 2017.

[24] Kollias D, Nicolaou M A, Kotsia I, et al. Recognition of Affect in the Wild Using Deep Neural Networks[Z]. 2017: 1972-1979.