第 58 卷 第 2 期/2021 年 1 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于嵌入注意力机制层级 LSTM 的音视频情感识别

刘天宝，张凌涛*，于文涛，魏东川，范轶军

中南林业科技大学计算机与信息工程学院，湖南 长沙 410004

摘要 对于语音的情感识别，针对单层长短期记忆(LSTM)网络在解决复杂问题时的泛化能力不足，提出一种嵌入自注意力机制的堆叠 LSTM 模型，并引入惩罚项来提升网络性能。对于视频序列的情感识别，引入注意力机制，根据每个视频帧所包含情感信息的多少为其分配权重后再进行分类。最后利用加权决策融合方法融合表情和语音信号，实现最终的情感识别。实验结果表明，与单模态情感识别相比，所提方法在所选数据集上的识别准确率提升 4%左右，具有较好的识别结果。

关键词 图像处理；情感识别；全卷积神经网络；长短期记忆网络；注意力机制；多模态融合

中图分类号 TP302.1 **文献标志码** A **doi**：10.3788/LOP202158.0210017

Hierarchical LSTM-Based Audio and Video Emotion Recognition With Embedded Attention Mechanism

Liu Tianbao, Zhang Lingtao*, Yu Wentao, Wei Dongchuan, Fan Yijun

College of Computer and Information Engineering, Central South University of Forestry and Technology, Changsha, Hunan 410004, *China*

Abstract A single-layer long short term memory (LSTM) network is not generalizable to solve complex speech emotion recognition problems. Therefore, a hierarchical LSTM model with a self-attention mechanism is proposed. Penalty items are introduced to improve network performance. For the emotion recognition of video sequences, the attention mechanism is introduced to assign a weight to each video frame according to its emotional information and then classify these frames. The weighted decision fusion method is used to fuse expressions and speech signals to achieve the final emotion recognition. The experimental results demonstrate that compared with single-modal emotion recognition, the recognition accuracy of the proposed method on the selected data is improved by approximately 4%, thus the proposed method has a better recognition results.

Key words image processing; emotion recognition; fully convolutional neural network; long short term memory network; attention mechanism; multimodal fusion

OCIS codes 100.0100; 100.2960; 100.4996

1 引 言

目前，由于人工智能技术的飞速发展，人类已不单单通过单纯的用户指令来进行人机交互。情感识别在其中的地位日益提升，其中语音和面部表情的识别是情感识别应用领域的关键组成部分之一，例如自动驾驶汽车、智能电话语音助手、人类心理分析及医疗服务等。随着研究人员的不断探究，对情感的识别也逐渐从原来的单模态向双模态甚至多模态转变，相比于单模态的情感识别，多模态往往有着更好的识别率和鲁棒性[1-3]。

语音情感识别的目的是从原始的语音信号中提

收稿日期：2020-07-06；**修回日期**：2020-08-14；**录用日期**：2020-09-08

基金项目：国家自然科学基金(61602529)

* **E-mail**：158809488@qq.com

取某些可用特征，如声道频谱、韵律及其他非线性特征，然后对其进行情感状态的识别和分类。传统的语音情感识别技术包括隐马尔可夫模型（HMM）、人工神经网络（ANN）、高斯混合模型（GMM）、支持向量机（SVM）及 K 近邻（KNN）等。Nwe 等[4]将 HMM 作为分类器，证明当 HMM 为 4 状态时，它用于情感识别分类时的识别效果最好。由于深度学习的发展和普及，许多研究人员将深度神经网络和语音情感识别结合起来，并取得了很好的实验效果。Satt 等[5]使用卷积神经网络（CNN）和长短期记忆（LSTM）网络对语音频谱进行处理，取得了良好的识别效果。由于语音的连贯性和所提取的情感信息有着一定的上下文关系，故本文构建一个基础 CNN-LSTM 架构对音频中的语音频谱进行建模。单层 LSTM 网络在处理复杂问题时有时会出现表达能力较差的情况，Sutskever 等[6]构建了 4 层 LSTM 网络并将它应用在编码器-解码器，实现了良好的实验效果，其提取出的深层次模型要明显优于浅层次模型。Irsoy 等[7]提出的具有多层体系架构的循环神经网络（RNN）在观点挖掘领域要明显优于传统的浅层 RNN。因此本文构建多层 LSTM 网络，通过利用多层 LSTM 网络提取每帧频谱的抽象信息来获得更好的性能。受很多动物和人在接受视觉信息时会聚集在某些区域而不是整个画面的这种视觉注意力的启发，研究人员将注意力机制引入深度学习中，这样不仅提高了识别的准确性而且很好地降低了信息处理的工作量。目前该技术被广泛应用于机器翻译、语音识别等领域。Lin 等[8]在基于文本的机器翻译方法中应用了一种新的注意力模型，与以往不同的是，该模型可以通过自己的信息来更新迭代模型的参数，该方法能够更好地聚焦句子的重要部分。Guo 等[9]提出一种具有可以并行处理多图像学习功能的新型 CNN 深度架构。研究人员通过最佳融合各种输入信号、特征、情感语音来进行广泛的实验。张石清等[10]提出了一种对监督局部线性嵌入策略进行某些改进，然后将其应用到情绪特征的降维方法。本文基于 CNN-LSTM 的语音情感识别网络基线模型，提出一种在各层 LSTM 网络间引入自注意力机制的改进方法，为各层 LSTM 网络的隐藏状态和单元状态分配权重，这样可以使不同层级得到更明显的区分，表示不同特征层级的关系，提升所提取特征的非线性表达能力。

传统面部表情识别方法流程为：首先从输入的数据集图片中检测出人脸图片和五官等关键部分，然后从中提取关于面部表情的特征，如几何特征、深度特征等，最后对这些特征进行训练和分类。由于所用到数据集中的视频并非每一帧都与情感表达相关，因此本文使用注意力机制，根据每个视频帧所包含情感信息的多少来为其分配权重。对于大多数情感识别的工作，往往通过对两个或多个模态信号进行融合来提升系统的识别效果[11]。一般来说，不同模态信号的融合可以自上而下地归为 3 类：判决层融合、中间层融合及特征层融合[12]。本文通过利用加权融合的策略来对面部图像和语音信号进行判决层融合，与融合前的各系统相比，该方法有着更好的识别率。

综上所述，对于音频的情感识别，本文基于 CNN-LSTM 基线模型，提出一种嵌入注意力机制的层级 LSTM 模型。所提方法首先通过 CNN 提取语音信号的特征并将其送入 LSTM 网络；再在层级 LSTM 间引入自注意力机制，使每层 LSTM 得到更好的区分，增强系统的鲁棒性；最后在损失函数中引入惩罚项，使网络性能得到进一步提升，这样可以使各个层的状态向量更加多样化，最终得到分类结果。实验结果表明，所提方法有着良好的识别效果。对于视频的情感识别，本文采用基于深度卷积网络（VGGNet）模型的改进模型等一系列方法来提取视频序列中每帧图像特征；然后通过注意力机制为每帧图像分配权重，以此筛选出与情感信息最相关的图像帧；此后进行情感识别；最后在决策层利用加权融合方法对两个模态进行融合，系统性能得到进一步提升。

2 基于嵌入注意力机制层级 LSTM 的音视频情感识别方法

将语音信息和面部表情结合在一起以实现情感识别。所提方法流程如图 1 所示。

2.1 语音情感识别方法

基于深度卷积神经网络的体系结构在图像建模和视频建模（例如图像的识别、分类、检索）中取得了优异的性能。为了实现语音情感识别，将语音信号转换为图像后再进行分类工作[13]，将 CNN 和 LSTM 网络无缝结合以实现语音情感识别。Shelhamer 等[14]针对图像的语义分割，提出一种基于全卷积网络的工作方法，该方法有效解决了网络输入大小的局限问题。故本文去掉 AlexNet 中的全连接层来构建一个全卷积网络用于特征提取，由于情感信息是存在于图像中的高级信息，故用 CNN

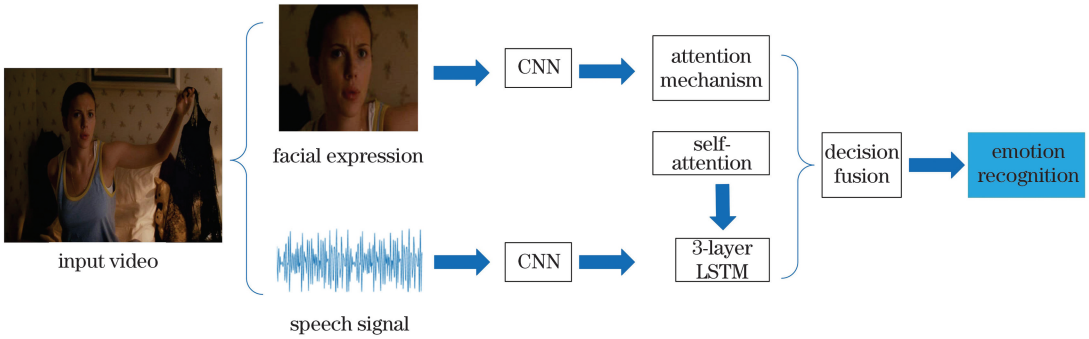

图 1　音视频情感识别系统流程图

Fig. 1　Flow chart of audio and video emotion recognition system

可以更好地捕获到此类信息及其抽象特征。常规 RNN 由于梯度消失和梯度爆炸问题，难以从随机长度输入序列中学习特征；LSTM 是一种特殊类型的 RNN，它能够学习长期依赖的信息，适用于处理和预测时间序列中间隔和延迟很长的重要事件，故用 LSTM 解决这个问题[15]。标准的递归神经元定义为

$$\boldsymbol{h}_i^{d+1}=f(\boldsymbol{a}_i^{d+1}),\tag{1}$$

$$\boldsymbol{a}_i^{d+1}=\sum_j \boldsymbol{w}_{ij}\boldsymbol{x}_j^{d+1}+\sum_k \boldsymbol{u}_{ik}\boldsymbol{h}_k^{d},\tag{2}$$

式中：函数 $f(\cdot)$ 为非线性激活函数；$\boldsymbol{h}_i^d$ 为第 d 层的第 i 个神经元的状态；$\boldsymbol{x}$ 为先前层的神经元；$\boldsymbol{w}$ 和 $\boldsymbol{u}$ 均为连接权重；j 为第 $d+1$ 层的神经元个数；k 为第 d 层的神经元个数。递归神经元如图 2 所示，其中 $\boldsymbol{c}_t$ 表示当前神经元的状态；t 为时间。考虑到三种类型的门，(2)式可以转换为

$$\boldsymbol{a}_i^{d+1}=\boldsymbol{c}_i^{d+1}\boldsymbol{a}_i^{d}+\boldsymbol{b}_i^{d+1}g\left(\sum_j \boldsymbol{w}_{ij}\boldsymbol{x}_j^{d+1}+\sum_k \boldsymbol{u}_{ik}\boldsymbol{h}_k^{d}\right),\tag{3}$$

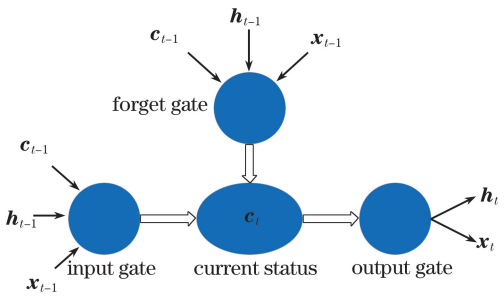

图 2　递归神经元结构

Fig. 2　Structure of recursive neuron

式中：符号 $\boldsymbol{c}$ 和 $\boldsymbol{b}$ 分别为遗忘门和输入门；与函数 $f(\cdot)$ 相似，$g(\cdot)$ 也为非线性激活函数。其输入包括三个分量，即来自上一层的信号、宿主单元、先前的输出，可以表述为

$$\boldsymbol{a}_i^{d+1}=g\left(\sum_j \boldsymbol{w}_{ij}^{\alpha}\boldsymbol{x}_j^{d+1}+\sum_k \boldsymbol{v}_{ik}^{\alpha}\boldsymbol{h}_k^{d}+\boldsymbol{v}_i^{\alpha}\boldsymbol{a}_i^{d+i_\alpha}\right),\tag{4}$$

式中：α 为门的范围；$\boldsymbol{v}$ 为宿主单元的连接权重。

Bahdanau 等[16]在机器翻译领域应用注意力机制，将其引入到模型中的输入和输出之间，从而使模型的性能得以提升。注意力机制的主要工作原理：将 source 中的元素想象成由一系列〈$\boldsymbol{K}$，$\boldsymbol{V}$〉数据对构成，给定元素 $\boldsymbol{Q}$，通过计算 $\boldsymbol{Q}$ 和每个 $\boldsymbol{K}$ 的相关性，得到每个 $\boldsymbol{K}$ 对应 $\boldsymbol{V}$ 的权重系数，然后对 $\boldsymbol{V}$ 进行加权求和，得到最终的 attention 数值 $\boldsymbol{Y}_{\mathrm{att}}$，如图 3 所示。$\boldsymbol{Y}_{\mathrm{att}}$ 可以表述为

$$\boldsymbol{Y}_{\mathrm{att}}=\sum_{z=1}^{l}\mathrm{Similarity}(\boldsymbol{Q},\boldsymbol{K}_z)\times\boldsymbol{V}_z,\tag{5}$$

式中：l 为 source 的长度。自注意力机制并不是指 target 和 source 之间的 attention 机制，而是 source 或 target 内部元素之间发生的注意力机制，可以理解为 $\boldsymbol{K}=\boldsymbol{V}=\boldsymbol{Q}$ 的情形。自注意力机制能够更容易捕获输入序列中长距离相互依赖的特征。

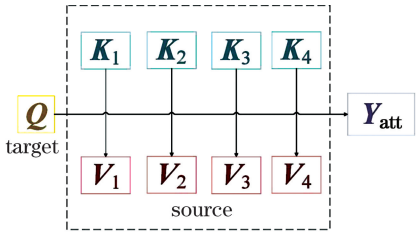

图 3　注意力机制原理图

Fig. 3　Schematic of attention mechanism

在所应用的栈式 LSTM 网络中，将三个 LSTM 堆叠在一起，该模型可以学习更高层次的时域特征表示。利用 LSTM 对序列数据进行操作，这意味着层的添加增加了输入观察时间的抽象级别，有着更好的表达能力。

为了使栈式 LSTM 网络中的各层 LSTM 有不同的占比，从而让所提网络性能得到进一步提升，在每层 LSTM 网络之间引入自注意力机制，与注意力机制不同的是，它可以通过自身的信息来进行迭代

更新。该部分方法流程如图 4 所示，此网络模型主要由嵌入自注意力机制的栈式 LSTM 网络构成，将栈式 LSTM 的隐藏状态和单元状态作为自注意力机制模块的输入，所输出的是相应的权重向量。

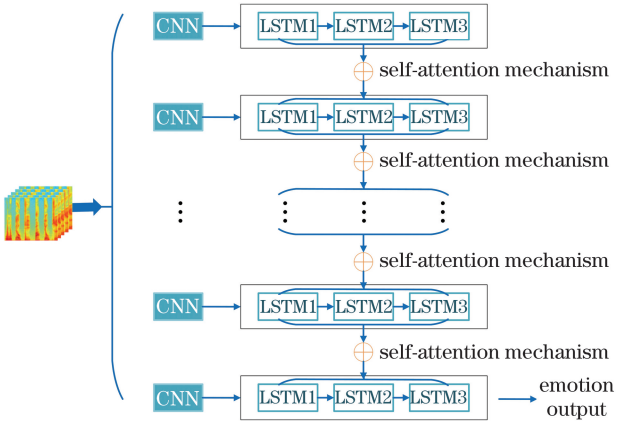

图 4　嵌入注意力机制的栈式 LSTM 模型示意图

Fig. 4　Schematic of stacking LSTM model with attention mechanism

$$\boldsymbol{u}^t=\boldsymbol{v}^{\mathrm{T}}\tanh(\boldsymbol{W}\boldsymbol{X}_t+\boldsymbol{b}),\tag{6}$$

$$\boldsymbol{a}^t=\mathrm{Softmax}(\boldsymbol{u}^t),\tag{7}$$

式中：向量 $\boldsymbol{X}_t$ 的维度为 $n\times r$；向量 $\boldsymbol{W}$ 的维度为 $r\times d_{\mathrm{a}}$；$\boldsymbol{b}$，$\boldsymbol{v}^{\mathrm{T}}$ 均为维度为 d_{a} 的向量；$\boldsymbol{W}$，$\boldsymbol{b}$，$\boldsymbol{v}^{\mathrm{T}}$ 均为网络模型的参数。$\boldsymbol{X}_t$ 是自注意力机制模块的输入，代表栈式 LSTM 中某一层的隐藏状态 $\boldsymbol{Y}_t$ 或单元状态 $\boldsymbol{Z}_t$。

$$\boldsymbol{Y}_t=(\boldsymbol{y}_t^{(1)},\boldsymbol{y}_t^{(2)},\cdots,\boldsymbol{y}_t^{(l)}),\tag{8}$$

$$\boldsymbol{Z}_t=(\boldsymbol{z}_t^{(1)},\boldsymbol{z}_t^{(2)},\cdots,\boldsymbol{z}_t^{(l)})。\tag{9}$$

接下来对权重向量 $\boldsymbol{a}^t$ 与 LSTM 的状态值进行乘积，即

$$\boldsymbol{G}_t=\boldsymbol{a}^t\boldsymbol{X}_t,\tag{10}$$

式中：$\boldsymbol{G}_t$ 为栈式 LSTM 经过更新后所得到的加权向量 $\boldsymbol{Y}'_t$ 或 $\boldsymbol{Z}'_t$。自注意力机制计算后，可以根据重要性对栈式 LSTM 中的各层网络分配不同的权重，网络得到了一定程度的优化并提升了层级特征的表达能力。

由于相邻时间步之间的自注意力机制往往会分配相近的权重，可以在其中加入惩罚项来防止该问题的发生，使不同层级状态权重向量更具有多样性。惩罚项在优化权重的同时，不仅减少了多余的特征信息，而且使栈式 LSTM 中的层级关系更具差异化，所以采用统计方差的方法来对网络进行优化。

$$\boldsymbol{P}=\frac{1}{T}\sum_{t}^{T}\sum_{i}^{L}[(\boldsymbol{a}_{ti}-\boldsymbol{\mu})^2+(\boldsymbol{\beta}_{ti}-\boldsymbol{\eta})^2],\tag{11}$$

$$\boldsymbol{\mu}=\frac{1}{L}\sum_{i}^{L}\boldsymbol{\alpha}_{ti},\ \boldsymbol{\eta}=\frac{1}{L}\sum_{i}^{L}\boldsymbol{\beta}_{ti},\tag{12}$$

式中：$\boldsymbol{P}$ 为惩罚项；$\boldsymbol{\alpha}_{ti}$ 和 $\boldsymbol{\beta}_{ti}$ 分别为隐藏状态和单元状态在不同时间步和层级上的注意力权重。将 $\boldsymbol{P}$ 与原损失函数一起最小化对网络权值进行优化。原损失函数表达式为

$$\boldsymbol{L}_{\mathrm{d}}=-\log[p(\boldsymbol{y}\mid\boldsymbol{a})])-\boldsymbol{P},\tag{13}$$

式中：$-\log[p(\boldsymbol{y}\mid\boldsymbol{a})]$ 为交叉熵损失函数；$\boldsymbol{a}$ 为模型的实际输出；$\boldsymbol{y}$ 为样本标签。

2.2　视频情感识别方法

与传统手工提取人脸面部特征不同的是，深度神经网络可以直接从所输入的原始图片中提取所需特征。对于视频的图像分类，LSTM 网络通常被用来提取分析与图像有关的时间域信息对。Fan 等[17-18]为处理视频中时间域信息，引入三维卷积网络和 LSTM 网络。不过对于 AFEW 这种非标准的数据集来说，并不是所有的视频帧都与情感表达有关，对此，使用注意力机制来筛选出最富有情感信息的视频帧，以进行视频序列的情感识别分类工作。

所使用的系统如图 5 所示。首先把数据集的每一帧原始图片送入 CNN 进行特征提取，得到相关的特征；然后经注意力机制，根据特征序列中的向量与情感的相关程度来分配相应权重；之后对这些权重进行加权求和，得到情感向量；最后对其进行分类。

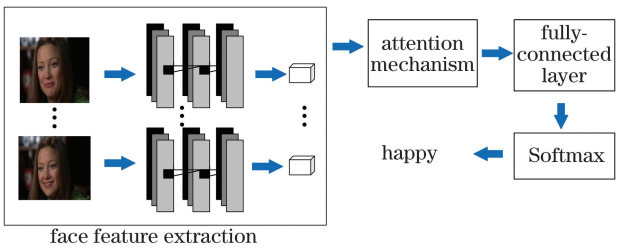

图 5　视频情感识别系统图

Fig. 5　Diagram of video emotion recognition system

采用文献[19]中提出的 4 种人脸特征提取方法来进行分类任务，特征提取的方法分别为 VGG-Face、FR-Net-A、FR-Net-B、FR-Net-C。本文分别将所提取的视频序列特征称为 EF-VGG、EF-A、EF-B、EF-C。

为了根据数据集视频中每一帧图片所包含情感信息的多少来为其分配不同的权重，采用注意力机制来计算每一帧图片的情感相关程度，然后对各帧图片进行加权融合得到情感特征。进行训练之前，需要对其所得特征进行降维，以降低计算复杂度，具体步骤表达式为

$$\tilde{\boldsymbol{v}}_c = \boldsymbol{W}\boldsymbol{v}_c + \boldsymbol{b}, \tag{14}$$

$$\tilde{\boldsymbol{e}}_c = \tilde{\boldsymbol{u}}^{\mathrm{T}} \tanh \tilde{\boldsymbol{v}}_c, \tag{15}$$

$$\tilde{\boldsymbol{\alpha}}_c = \frac{\exp(\lambda \tilde{\boldsymbol{e}}_c)}{\sum_{k=1}^{L} \exp(\lambda \tilde{\boldsymbol{e}}_k)}, \tag{16}$$

$$\tilde{\boldsymbol{v}} = \sum_{c=1}^{Q} \tilde{\boldsymbol{\alpha}}_c \tilde{\boldsymbol{v}}_c, \tag{17}$$

式中：$\tilde{\boldsymbol{v}}_c$ 为第 c 帧经过降维之后所得到的新特征向量；$\tilde{\boldsymbol{v}}$ 为情感特征向量；λ 为注意力机制所分配权重的大小。

2.3 音视频融合的情感识别方法

本文集成面部表情和语音信号来进行情感识别，而且利用决策融合方法来解决两个不同模态的融合问题。决策融合的目的是处理每种模型产生的类别，并利用特定的标准再进行重新区分。人脸面部表情识别和语音情感识别都使用 Softmax 函数来进行分类。将面部表情识别和语音情感识别的输出分别定义为

$$\boldsymbol{S}^{\mathrm{face}} = \{\boldsymbol{S}_1^{\mathrm{face}}, \boldsymbol{S}_2^{\mathrm{face}}, \boldsymbol{S}_3^{\mathrm{face}}, \boldsymbol{S}_m^{\mathrm{face}}\}, \tag{18}$$

$$\boldsymbol{S}^{\mathrm{speech}} = \{\boldsymbol{S}_1^{\mathrm{speech}}, \boldsymbol{S}_2^{\mathrm{speech}}, \boldsymbol{S}_3^{\mathrm{speech}}, \cdots, \boldsymbol{S}_m^{\mathrm{speech}}\}, \tag{19}$$

式中：m 为情感类别的数量。加权决策融合计算为

$$\boldsymbol{S} = w_0 \boldsymbol{S}^{\mathrm{face}} + w_1 \boldsymbol{S}^{\mathrm{speech}}, \ w_0 + w_1 = 1, \tag{20}$$

式中：w_0 和 w_1 分别为两个模态所分配的权重。

3 实验结果与分析

对所提方法进行有效性验证。实验是在 NIVADA 1060ti，4.00GB RAM 的计算机上进行的。实验中使用三个人类情感数据集，包括 RML[20]、AFEW6.0[21]及 eNTERFACE′05[22]。

3.1 数据集简介

RML：RML 是一个双模态数据集，其中包括面部表情和语音信息。该数据集由 720 个视频样本组成，其中包含 6 种基本表情。数据集的采样率为 44100 Hz，视频帧率为 30 frame/s。

AFEW6.0：该数据集由 773 个训练样本、383 个验证样本及 593 个测试样本组成，总共包含 7 类情感。该数据集的数据均是从电影、电视、脱口秀等影视片段中采集到的，更能体现真实场景下的人类情感表达。

eNTERFACE′05：该数据集将 42 个具有不同国籍的个体作为视频样本，其中包括 1263 个视频。在这些视频剪辑中，有 81%是从男性那里收集的，而 19%是从女性那里收集的。每帧的尺寸为 720×576。

在本实验中，首先利用旋转、翻转、颜色失真、镜像变换操作对数据进行扩充。使用 Caffe 深度学习框架来实现所提方法。最初对整个数据集进行 100 个周期的训练，批次大小为 32；初始学习率为 0.01，在 10000 次迭代后将其设置为 0.005；将权重衰减量和动量分别设置为 0.0002 和 0.9。深度情感识别模型采用随机梯度下降(SGD)方案训练。

3.2 语音情感识别

由于传统语音数据的规模小，将具有不同权重和信噪比的高斯白噪声集成到原始语音信号上。最初将训练周期设置为 100，批次大小为 32。初始学习率为 0.01，在 10000 次迭代后将学习率设置为 0.005。将权重衰减量和动量分别设置为 0.00001 和 0.9。通过随机梯度下降(SGD)算法来训练深层模型。长期和短期存储网络中的隐藏层单元数固定为 128。比较结果如表 1 所示。

表 1 语音情感识别实验的识别率对比

Table 1 Comparison of recognition rate in speech emotion recognition experiment

Network	RML	AFEW6.0	eNTERFACE′05
SVM[23]	0.6020	0.3790	0.4831
Random forest[24]	0.6528	0.3508	0.4711
LSTM+CNN[25]	0.8546		0.4915
CNN	0.8363		0.4691
CNN+LSTM	0.8446	0.4217	0.4952
Proposed network	0.9011	0.5473	0.5932

3.2.1 LSTM 堆叠层数对系统识别率的影响

为了探究 LSTM 层数是否会对实验结果有相应的改进作用，基于基线模型，在不同层数的 LSTM 下进行对比实验。

图 6 为不同层数的 LSTM 对系统识别率影响的结果。实验结果表明，相比于单层网络，多层的 LSTM 有

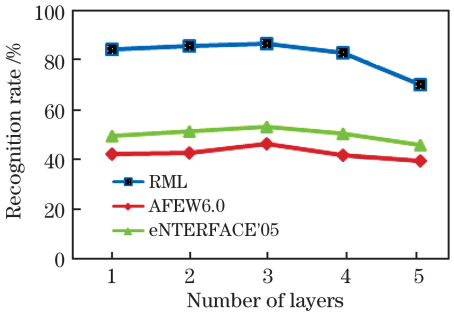

图 6 LSTM 层数和识别率的关系

Fig. 6 Relationship between LSTM layers and recognition rate

着更好的识别效果，可以更好地提取序列中的抽象特征。当层数为 3 时，在所选取数据集上的识别效果达到最佳；当层数大于 3 时，所呈现的识别效果逐渐下降，这是因为层数过多，梯度变小，浅层的 LSTM 权重不能得到更新。综上，选取 LSTM 层数为 3。

3.2.2　层级注意力机制对系统识别率的影响

系统引入注意力机制后可以使网络的不同层级在各个时间步上有选择的被关注。为了研究注意力机制是否对网络的提升有一定影响，设计了相应的对比实验，结果如表 2 所示，结果说明，注意力机制的引入使模型的识别效果得到了一定的提升。给具有注意力机制的栈式 LSTM 中的各层分配不同占比的方法有利于网络筛选出更有用的信息，提升模型的层级表达能力，并更有利于提取图像的抽象特征。实验结果表明，注意力机制可以很好地改善识别效果，系统识别率提升大致 3%。

表 2　层级注意力机制的识别率对比

Table 2　Recognition rate comparison of hierarchical attention mechanism

Dataset	3-layer LSTM	
	Ordinary	Add attention mechanism
RML	0.8661	0.8873
AFEW6.0	0.4633	0.4965
eNTERFACE′05	0.5315	0.5739

3.2.3　惩罚项对系统识别率的影响

在注意力机制中，通过改变注意力的权重系数来达到改善识别效果的目的。其中，惩罚项可以用来更新权重系数，不同的权重系数得到的识别模型有所差异。通过引入方差，得到不同权重系数情况下的区别，再使用反向传播算法使方差最大化。表 3 是加入惩罚项与不加惩罚项情况下，三个数据集识别率的对比情况。从表 3 可以看出，添加惩罚项能提高不同数据集下的网络性能。

表 3　惩罚项情况下的识别率对比

Table 3　Recognition rate comparison under penalty items

Dataset	Ordinary	Add penalty
RML	0.8873	0.9011
AFEW6.0	0.4965	0.5473
eNTERFACE′05	0.5739	0.5932

3.3　面部表情识别

分别在 AFEW 6.0、RML、eNTERFACE′05 数据集上对所设计的模型进行实验验证。在所提模型中，所提取的特征长度为 256，训练过程中，将 batch size 设置为 64，迭代 200 个 epoch，学习率为 3×10^{-5}。分别使用 EF-A、EF-B、EF-C、EF-VGG 作为系统输入的特征，由于所提取的特征之间存在一定程度上的差异，在输入系统前要对其进行规整，表达式为

$$f=\frac{f-f_{\alpha_{\text{train}}}}{f_{\nu_{\text{train}}}},\tag{21}$$

式中：$f_{\alpha_{\text{train}}}$ 为训练集的特征均值；$f_{\nu_{\text{train}}}$ 为训练集的特征方差。此过程的目的是防止特征的幅度波动过大，加速系统模型的训练。实验结果如表 4 所示。经过研究发现，视频中情感的表达和时序性关联并不明显，甚至可以倒序播放，真正起作用的是视频序列中包含有情感信息的视频帧，实验结果表明，注意力机制的引入有着较好的实验效果。所以采用 FR-Net-B 网络来进行面部表情的特征提取。

表 4　面部表情的识别率

Table 4　Recognition rate of facial expression

Video sequence feature	RML	AFEW6.0	eNTERFACE′05
EF-A	0.8653	0.5074	0.7458
EF-B	0.8812	0.5185	0.7974
EF-C	0.8232	0.4713	0.7515
EF-VGG	0.8346	0.4882	0.7627

3.4　音视频融合的情感识别

利用面部表情识别和语音情感识别来进行决策融合。对在决策层融合的两种模态信号而言，两种信号与情感表达程度大小各异，如何合理分配权重也成为融合的关键，由实验效果和经验可知，该实验可为视频信号分配更大的权重。表 5 为三个数据集的权重。图 7 显示了与其他特征融合方法如 Denseface-Net[11]、FBP[26] 相比的识别效果。由图 7 可以看出：基于双模态融合的情感识别优于单模态情感识别；通过融合各种面部表情，所提模型可以将整体语音识别率提高大约 4%，而且在三个数据集上均取得了良好的效果。同时所提识别模型具有一定的普遍性[25,27]。在本工作中，尽管加权决策融合与特征融合相比有一定的局限性，但实验结果却更好。

表 5　三种数据集的权重设置

Table 5　Weight settings on three datasets

Dataset	Facial expression recognition	Speech expression recognition
RML	0.60	0.40
AFEW6.0	0.75	0.25
eNTERFACE′05	0.80	0.20

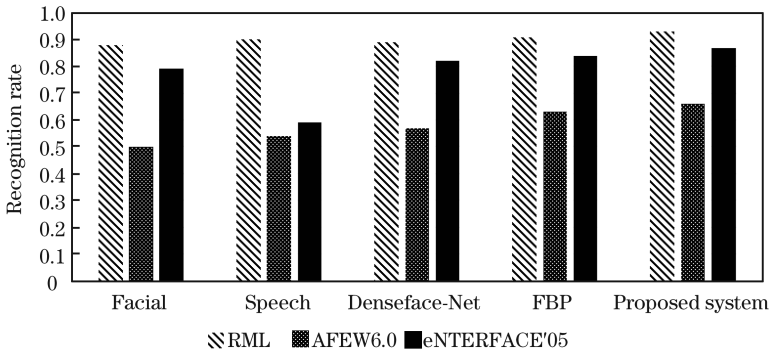

图 7 不同特征融合算法的性能比较

Fig. 7 Performance comparison of different feature fusion algorithms

4 结 论

提出一种应用于语音情感识别的基于自注意力机制的堆叠 LSTM 模型。与其他的现有模型相比，它能提取时间序列数据中更复杂的依赖关系，并且通过注意力机制，可以使模型聚焦于更为重要的层级，在引入惩罚项后，不同时间步之间的注意力向量更具有多样性。针对视频情感识别，引入了注意力机制，相比其他现有的模型而言，大大减少了模型训练学习的时间，同时还具有良好的识别效果。最后应用加权融合算法，将面部信号和语音信号融合在一起，融合后的系统识别率相较于其他系统提升 4%左右，进一步提升了识别的准确率。除了语音和面部表情，人类的情感还有着多种载体，比如姿态动作、生理信号等，如何更加有效利用并融合多模态的信息来提升识别准确率将是下一步的研究工作。

参 考 文 献

[1] Yuan P P, Zhang L. Pedestrian attribute recognition based on deep learning[J]. Laser & Optoelectronics Progress, 2020, 57(6): 061001.
袁配配, 张良. 基于深度学习的行人属性识别[J]. 激光与光电子学进展, 2020, 57(6): 061001.

[2] Liu F, Li M J, Hu J W, et al. Expression recognition based on low pixel face images [J]. Laser & Optoelectronics Progress, 2020, 57(10): 101008.
刘芾, 李茂军, 胡建文, 等. 基于低像素人脸图像的表情识别[J]. 激光与光电子学进展, 2020, 57(10): 101008.

[3] Zhang Y C, Sun Z W. Identity authentication for smart phones based on an optimized convolutional deep belief network [J]. Laser & Optoelectronics Progress, 2020, 57(8): 081009.
张义超, 孙子文. 基于优化卷积深度信念网络的智能手机身份认证方法[J]. 激光与光电子学进展, 2020, 57(8): 081009.

[4] Nwe T L, Foo S W, de Silva L C. Speech emotion recognition using hidden Markov models[J]. Speech Communication, 2003, 41(4): 603-623.

[5] Satt A, Rozenberg S, Hoory R. Efficient emotion recognition from speech using deep learning on spectrograms[J]. Proceedings of Interspeech 2017, 2017: 1089-1093.

[6] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks [C] // Proceedings of the 27th International Conference on Neural Information Processing Systems, December 8-13, 2014, Montreal, Quebec, Canada. New York: Curran Associates, 2014, 2: 3104-3112.

[7] Irsoy O, Cardie C. Deep recursive neural networks for compositionality in language[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems, December 8-13, 2014, Montreal, Quebec, Canada. New York: Curran Associates, 2014, 2: 2096-2104.

[8] Lin Z H, Feng M W, dos Santos C N, et al. A structured self-attentive sentence embedding [EB/OL]. (2017-03-09)[2020-07-05]. https://arxiv.org/abs/1703.03130.

[9] Guo Z H, Zhang L, Zhang D. A completed modeling of local binary pattern operator for texture classification [J]. IEEE Transactions on Image Processing, 2010, 19(6): 1657-1663.

[10] Zhang S Q, Li L M, Zhao Z J. Speech emotion recognition based on an improved supervised manifold learning algorithm [J]. Journal of Electronics & Information Technology, 2010, 32(11): 2724-2729.
张石清, 李乐民, 赵知劲. 基于一种改进的监督流形学习算法的语音情感识别[J]. 电子与信息学报, 2010, 32(11): 2724-2729.

[11] Wang S, Wang W X, Zhao J M, et al. Emotion recognition with multimodal features and temporal models [C] // Proceedings of the 19th ACM International Conference on Multimodal Interaction-

ICMI 2017, November 3-17, 2017, Glasgow, UK. New York: ACM Press, 2017: 598-602.

[12] Wu C H, Lin J C, Wei W L. Survey on audiovisual emotion recognition: databases, features, and data fusion strategies[J]. APSIPA Transactions on Signal and Information Processing, 2014, 3: e12.

[13] Abdel-Hamid O, Mohamed A R, Jiang H, et al. Convolutional neural networks for speech recognition [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(10): 1533-1545.

[14] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.

[15] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[16] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [EB/OL]. (2016-05-19) [2020-07-05]. https://arxiv.org/abs/1409.0473.

[17] Fan Y, Lu X J, Li D, et al. Video-based emotion recognition using CNN-RNN and C3D hybrid networks [C] // Proceedings of the 18th ACM International Conference on Multimodal Interaction-ICMI 2016, October 31-November 16, 2016, Tokyo, Japan. New York: ACM Press, 2016: 445-450.

[18] Nguyen D, Nguyen K, Sridharan S, et al. Deep spatio-temporal features for multimodal emotion recognition[C] // 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), March 24-31, 2017, Santa Rosa, CA, USA. New York: IEEE Press, 2017: 1215-1223.

[19] Knyazev B, Shvetsov R, Efremova N, et al. Leveraging large face recognition data for emotion classification [C] // 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), May 15-19, 2018, Xi′an, China. New York: IEEE Press, 2018: 692-696.

[20] Wang Y J, Guan L. Recognizing human emotional state from audiovisual signals[J]. IEEE Transactions on Multimedia, 2008, 10(4): 659-668.

[21] Dhall A, Goecke R, Joshi J, et al. EmotiW 2016: video and group-level emotion recognition challenges [C] // Proceedings of the 18th ACM International Conference on Multimodal Interaction-ICMI 2016, October 31-November 16, 2016, Tokyo, Japan. New York: ACM Press, 2016: 427-432.

[22] Martin O, Kotsia I, Macq B, et al. The eNTERFACE′05 audio-visual emotion database[C] // 22nd International Conference on Data Engineering Workshops (ICDEW′06), April 3-7, 2006, Atlanta, GA, USA. New York: IEEE Press, 2006.

[23] Avots E, Sapiński T, Bachmann M, et al. Audiovisual emotion recognition in wild[J]. Machine Vision and Applications, 2019, 30(5): 975-985.

[24] Noroozi F, Marjanovic M, Njegus A, et al. Audio-visual emotion recognition in video clips[J]. IEEE Transactions on Affective Computing, 2017, 10(1): 60-75.

[25] Wang X S, Chen X, Cao C J. Human emotion recognition by optimally fusing facial expression and speech feature [J]. Signal Processing: Image Communication, 2020, 84: 115831.

[26] Zhang Y Y, Wang Z R, Du J. Deep fusion: an attention guided factorized bilinear pooling for audio-video emotion recognition [C] // 2019 International Joint Conference on Neural Networks (IJCNN), July 14-19, 2019, Budapest, Hungary. New York: IEEE Press, 2019.

[27] Dangol R, Alsadoon A, Prasad P W C, et al. Speech emotion recognition using convolutional neural network and long-short term memory [J]. Multimedia Tools and Applications, 2020, 79: 32917-32934.