基于原始波形的端到端阿尔茨海默症检测方法

陈旭初[1],张卫强[1],马 勇[2]
(1. 清华大学电子工程系,北京 100084;2. 江苏师范大学语言科学与艺术学院,江苏徐州 221009)

摘 要: 阿尔茨海默症(Alzheimer's Disease,AD)是一种退行性疾病,随着病情加重,患者的语言能力逐渐减弱. 目前已经有研究者使用梅尔谱图、梅尔频率倒谱系数(Mel Frequency Cepstral Coefficient,MFCC)等声学特征对阿尔茨海默症患者和健康人进行分类,但是对于使用神经网络从原始波形提取特征进行阿尔茨海默症检测还缺少进一步的探索. 本文提出一种基于原始波形的端到端阿尔茨海默症检测方法. 该方法使用一维卷积从原始波形中提取时间维度特征,并使用含有膨胀卷积的残差块提取更复杂的特征. 为进一步提高性能,在残差块中引入挤压-激励模块. 在全国人机语音通讯学术会议(National Conference on Man-Machine Speech Communication,NCMMSC)2021 AD 数据集上,本文提出的模型在长音频测试集、短音频测试集分别达到了 86.55% 和 81.35% 的准确率,比基线系统分别提高了 6.75%、7.35%. 在 INTERSPEECH2020 ADReSS 数据集上,模型的准确率为 66.67%,比基线系统提高 4.17%.

关键词: 阿尔茨海默症;语音检测;残差块;挤压-激励模块;端到端
基金项目: NSFC-通用技术基础研究联合重点基金(No.U1836219)
中图分类号: TP391.5 **文献标识码:** A **文章编号:** 0372-2112(2023)12-3582-09
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220162

Raw Waveform-Based End-to-End Alzheimer's Disease Detection Method

CHEN Xu-chu[1], ZHANG Wei-qiang[1], MA Yong[2]
(1. *Department of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*;
2. *School of Linguistic Sciences and Arts, Jiangsu Normal University, Xuzhou, Jiangsu* 221009, *China*)

Abstract: Alzheimer's disease (AD) is a degenerative disease, as the disease worsens, the patient's language ability gradually decreases. Some researchers have already used acoustic features such as Mel spectrogram and Mel frequency cepstral coefficient (MFCC) to classify AD patients and healthy individuals, but there is a lack of further exploration on using neural networks to extract features from raw waveforms for AD detection. In this paper, we propose an end-to-end AD detection method based on raw waveforms. The method uses one-dimensional convolution to extract time-dimensional features from the original waveform and uses a residual block containing an inflated convolution to extract more complex features. To further improve performance, the squeeze-and-excitation block is introduced into the residual block. On the national conference on man-machine speech communication (NCMMSC) 2021 AD dataset, the model proposed in this paper achieves 86.55% and 81.35% accuracy on the long audio test set and short audio test set, respectively, which is 6.75% and 7.35% better than the baseline system, respectively. On the INTERSPEECH2020 ADReSS dataset, the accuracy of the model is 66.67%, an improvement of 4.17% over the baseline system.

Key words: Alzheimer's disease; speech detection; residual blocks; squeeze-and-excitation block; end-to-end
Foundation Item(s): Joint Fund of the National Natural Science Foundation of China (NSFC) and the Fundamental Research of General Technology (No.U1836219)

1 引言

阿尔茨海默症(Alzheimer's Disease,AD)是一种退行性疾病[1],已有的治疗手段仅能维持或减缓患者认知能力衰退的速度,而不能逆转已经恶化的痴呆. 2010 年,全球阿尔茨海默症等痴呆症患者总数估计为 3 560 万人,预计到 2030 年患者人数达到 6 570 万人,给家庭

收稿日期:2022-02-10;修回日期:2022-07-30;责任编辑:梅志强

和社会带来沉重的负担[2]. 阿尔茨海默症早期的筛查方法主要是依赖神经心理学测验等特异性一般的检查,比如精神状态检查[3]和蒙特利尔认知评估[4],临床确诊依然依赖核磁共振成像、脑脊液蛋白标记物定量等成本较高的检测方法[5]. 这种疾病的最初症状之一是语言能力的恶化,随着患者病情的加重,患者的失语、表达困难等症状更加明显[2,6]. 因此,开发基于语音的阿尔兹海默症筛查手段成为了一个近年来比较火热的研究领域.

目前已经有研究者从语音中提取人工设计的声学特征和语言学特征,在半结构化语音任务数据集中区分健康人和阿尔茨海默症患者[7~12]. 文献[7]使用卷积循环神经网络(Convolutional Recurrent Neural Network, CRNN)作为特征序列生成器,从语音的对数滤波器组特征(Filter Banks, FBank)特征中识别出音节作为新的特征序列,然后使用循环神经网络(Recurrent Neural Network, RNN)对健康人和阿尔茨海默症患者进行分类. 文献[8]提取了健康人和阿尔茨海默症患者语音的梅尔频率倒谱系数(Mel Frequency Cepstral Coefficient, MFCC)、eGeMAPS(extended Geneva Minimalistic Acoustic Parameter Set)[13]等特征,并将语音转录为文本后,提取语言学特征,采取单独使用声学特征、语言学特征,以及两种特征融合等策略,使用逻辑回归模型进行分类.

虽然也有研究者使用原始波形进行阿尔茨海默症检测,但主要还是对语音的韵律特征进行分析. 文献[9]对阿尔茨海默症患者、轻度认知障碍患者(Mild Cognitive Impairment, MCI)和健康人(Health Control, HC)三种类别的语音进行分析时,通过提取语音段、静音段的时长以及语音中单词出现的时间间隔等信息,利用均值、中位数等统计特征对三类数据进行多次二分类任务. 文献[10]首先统计患者发声和停顿的时间和持续时间、说话速度等信息,而后计算时间间隔、说话速度的平均值、方差等统计特征,从而实现对阿尔茨海默症患者和健康人的分类.

随着机器学习的发展,人们能够使用端到端网络从原始波形中学习特定的任务特征,从而最大限度地减少人工对输入数据的处理,这种从原始波形中提取特征的方法已有研究者进行了尝试[14~16],然而,对于使用神经网络从原始波形中提取特征进行阿尔茨海默症检测还缺少进一步的探索.

受到应用于音乐自动标注领域的样本级卷积神经网络[17](sample-level Convolutional Neural Networks, sampleCNN)启发,我们使用sampleCNN模型直接从语音的原始波形中提取音频特征用于阿尔茨海默症检测. 该模型使用卷积核大小为3的一维卷积从原始音频上提取特征,通过堆叠小过滤器,在保留感受野的同时,有效地学习更复杂的特征[18]. 该模型在音乐自动标记上取得了很好的效果. 但是对卷积神经网络而言,当网络层数较多时,将会产生“退化”现象,因此,文献[19]提出残差结构,通过在多个卷积层之间增加短路机制改善网络的性能,该结构在图像分类等任务上有较好的表现. 而在神经网络中使用膨胀卷积,可以在不扩大卷积核的情况下增加模型感受野,获得不同尺度的信息[20]. 卷积神经网络虽然可以通过多个通道提取数据的深层次特征,但是通道之间的权重是相同的. 文献[21]提出将挤压-激励(Squeeze-and-Excitation, SE)模块嵌入到神经网络,可以提取特征通道间的信息,从而有效降低图像分类的错误率.

本文的研究重点是提出一种改进的sampleCNN方法,该方法将sampleCNN模型中的卷积块更换为残差块,并在残差块中使用膨胀卷积以增加感受野. 为提高模型的分类效果,在残差块中加入挤压-激励模块. 这种方法在NCMMSC(National Conference on Man-Machine Speech Communication) 2021 AD和INTERSPEECH2020 ADReSS数据集上取得了较好的效果.

2　系统结构

基于原始波形的阿尔茨海默症检测系统主要由数据准备、神经网络、众数决策3个部分组成,系统结构如图1所示.

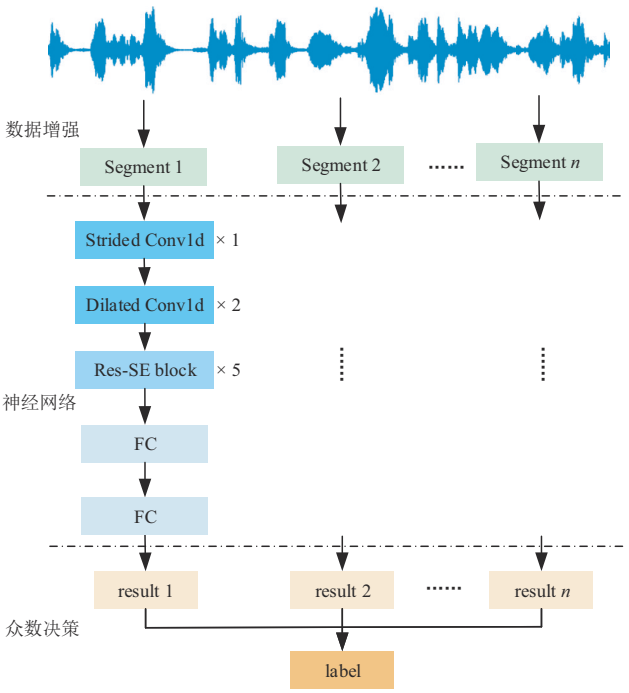


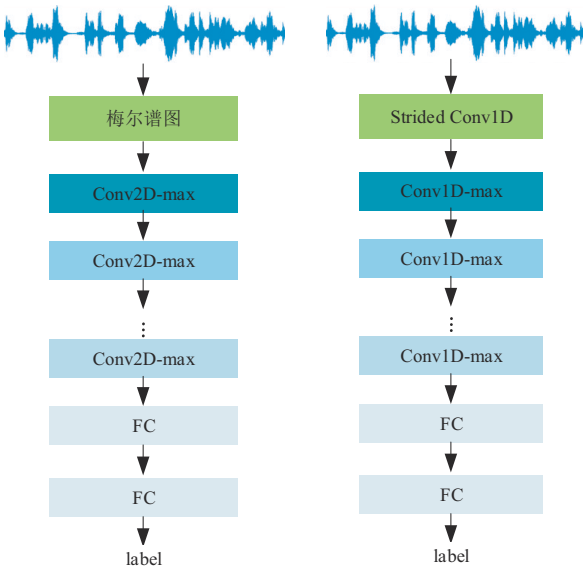
图1　Res-SE sampleCNN系统结构图

2.1 数据准备

由于通常情况下,数据集中患者语音的时长不同,为使神经网络的输入具有相同的数据长度,并且增加训练数据的数量,所以先对训练数据进行数据增强,将语音分割为多个时长相同的片段,并且相邻片段之间有一定的重叠. 在训练集中,赋予分割后的片段与原音频相同的标签.

2.2 基于原始波形的sampleCNN模型

本节主要对基础的sampleCNN模型进行介绍,并与基于梅尔谱图的卷积神经网络模型进行比较. 图2(*a*)、图2(*b*)分别为基于梅尔谱图的卷积神经网络模型和使用原始波形的sampleCNN模型的简化网络结构.



(*a*) 卷积神经网络模型 (*b*) sampleCNN模型

图2 输入为梅尔谱图的卷积神经网络模型和输入为原始波形的sampleCNN模型

梅尔谱图是手工设计的语音特征. 对于音频片段,通过分帧、加窗、短时傅里叶变换、线性到梅尔映射和幅度压缩等步骤,将时域信号转变为二维的时频表示,然后将梅尔谱图视为一个二维图像,作为神经网络的输入[22]. 如图2(*a*)所示,卷积神经网络使用多个二维卷积-最大池化模块(Conv2D-max)提取不同层次和尺度的特征,而后将其聚合为一个单一的特征向量,利用两个全连接层(FC layer)对标签进行预测. 需要通过多次实验,选择模型中使用的超参数,如窗长、窗移、梅尔滤波器组的个数等.

sampleCNN模型主要由一维卷积层和一维池化层组成[17]. 第一个一维卷积层滤波器的尺寸只有2个或3个样本点的长度,可以对原始波形进行小粒度采样. 图2(*b*)是以原始波形为输入的端到端sampleCNN模型的基础结构,采用1个一维卷积层和1个最大池化层(Conv1D-max)作为基本构建模块,卷积核大小为3,池化层大小也是3. 在第一个跨步卷积层之后,使用9个卷积-池化模块提取时间维度上的特征,最后使用全连接层进行预测.

2.3 Res-sampleCNN模型

为使用更深层的网络提取原始波形的特征,我们将sampleCNN中的卷积层替换为残差块. 文献[23]比较了不同结构的残差单元对性能的影响,认为将激活函数放在卷积层之前,形成预激活残差单元(pre-activation residual unit),此时网络更加容易优化.

图3(*a*)是Res-sampleCNN中使用的残差块的结构,我们使用膨胀卷积层作为残差结构中的第二个卷积层,以便在使用同样大小卷积核的情况下增加模型的感受野.

当上一层网络的输出为 $\boldsymbol{Y}\in\mathbf{R}^{C\times T}$ 时,残差块中第一层卷积的输出 $\boldsymbol{Y}_1\in\mathbf{R}^{C_1\times T_1}$ 为

$$\boldsymbol{Y}_1=\mathrm{StriedConv1D}\left(\sigma\left(\mathrm{BN}\left(\boldsymbol{Y}\right)\right)\right) \tag{1}$$

其中,C、T分别表示通道数、时间维度,$\mathrm{StriedConv1}D(\cdot)$表示图3(*a*)中的一维卷积,其卷积核大小$k$为3,步长$s$为2,$\sigma(\cdot)$表示ReLU函数,$\mathrm{BN}(\cdot)$表示批量归一化(BatchNormal)函数. 输出通道数C_1是输入通道数C的两倍,T_1为

$$T_1=\left\lfloor\frac{T-k}{s}\right\rfloor+1 \tag{2}$$

残差块中第二层卷积的输出 $\boldsymbol{Y}_2\in\mathbf{R}^{C_2\times T_2}$ 为

$$\boldsymbol{Y}_2=\mathrm{DilatedConv1D}\left(\sigma\left(\mathrm{BN}\left(\boldsymbol{Y}_1\right)\right)\right) \tag{3}$$

其中,$\mathrm{DilatedConv1}D(\cdot)$表示图3(*a*)中的膨胀卷积层,其卷积核大小$k$、步长$s$分别为3、1,膨胀率$d$为2,由此可以计算出感受野$F$:

$$F=k+(k-1)\times(d-1) \tag{4}$$

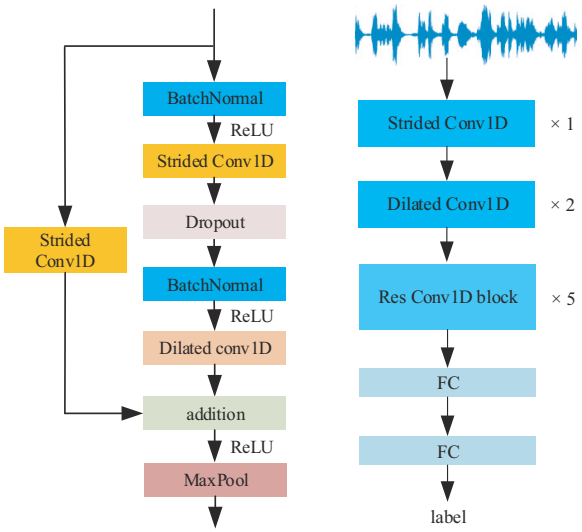
在卷积时通过补零保持维度不变. 此外,为了防止过拟合,在两个卷积层之间插入一个参数为0.1的dropout层.

图3(*b*)是我们提出的Res-sampleCNN网络结构. 对于输入的原始波形,首先经过1个卷积核大小为3、步长为3的一维跨步卷积层(StridedConv1D)来提取时间维度特征,输入通道数为1,输出通道数为64;然后使用两个卷积核大小为3、膨胀率为2的膨胀卷积层(DilatedConv1D)提取更大感受野的特征,输入和输出的通道数均为64;而后使用多个图3(*a*)所示的残差块提取更复杂的特征,除最后一个残差块的后面使用长度为4、步长为4的平均池化层进行下采样外,其它残差块的后面使用1个长度和步长都为2的最大池化层进行下采样. 网络最后使用两个全连接层进行预测.

最后一个全连接层的激活函数选择为sigmoid函数,并采用多分类交叉熵损失函数作为优化的目标函数:

$$\text{Loss}=-\sum_{i=0}^{n} y_i \log(p_i) \tag{5}$$

其中,$\boldsymbol{p}=[p_0,p_1,\cdots,p_n]$是概率分布,每个元素$p_i$表示样本属于第$i$类的概率,$\boldsymbol{y}=[y_0,y_1,\cdots,y_n]$是样本标签的独热编码.



(a) Res-Conv1D模块　　(b) Res-sampleCNN模型

图3　Res-Conv1D模块和Res-sampleCNN模型

当输入为长度为6 s、采样率为16 kHz的音频时,Res-SE sampleCNN模型各层的输出如表1所示.

2.4　Res-SE模块

为了进一步提高2.3节中残差块的学习能力,我们在其中加入挤压-激励(Squeeze-and-Excitation,SE)模块[21]. 挤压-激励模块由挤压(Squeeze)和激励(Excitation)两种操作组成,其结构如图4(a)所示,该模块可以使神经网络更加关注信息量大的通道特征,而抑制那些不重要的通道特征[24]. 我们将挤压-激励模块作为基本单元添加到图3(a)所示的残差块中,得到如图4(b)所示的Res-SE模块.

挤压-激励模块首先对卷积得到的特征图进行挤压操作,通过全局平均池化(GloAvgPool)得到通道的全局特征. 假设上一层网络的输出特征为$\boldsymbol{Y}\in\mathbf{R}^{C\times T}$,通过全局池化后,可将特征图经过挤压操作得到输出$\boldsymbol{z}\in\mathbf{R}^{C\times 1}$,$z$第$c$个元素为

$$z_c=\text{GloAvgPool}(\boldsymbol{u}_c)=\frac{1}{T}\sum_{i=1}^{T} u_{c,i} \tag{6}$$

其中,$\boldsymbol{u}_c$表示$\boldsymbol{Y}$的第c行.

然后,对$\boldsymbol{z}$进行激励操作,该操作是通过两个全连接层实现的,可以通过超参数α调整全连接层的维度,得到输出$\boldsymbol{s}\in\mathbf{R}^{C\times 1}$:

$$\boldsymbol{s}=\delta(\boldsymbol{W}_2,\sigma(\boldsymbol{W}_1,\boldsymbol{z})) \tag{7}$$

其中,$\sigma(\cdot)$和$\delta(\cdot)$分别表示ReLU激活函数和sigmoid函数,$\boldsymbol{W}_1$和$\boldsymbol{W}_2$表示全连接层. 第二个全连接层的输出经过sigmoid函数激活后,网络学习到通道之间的信息并调整输出权重的大小,将此权重与对应的原特征通道相乘便可得到加权后的特征.

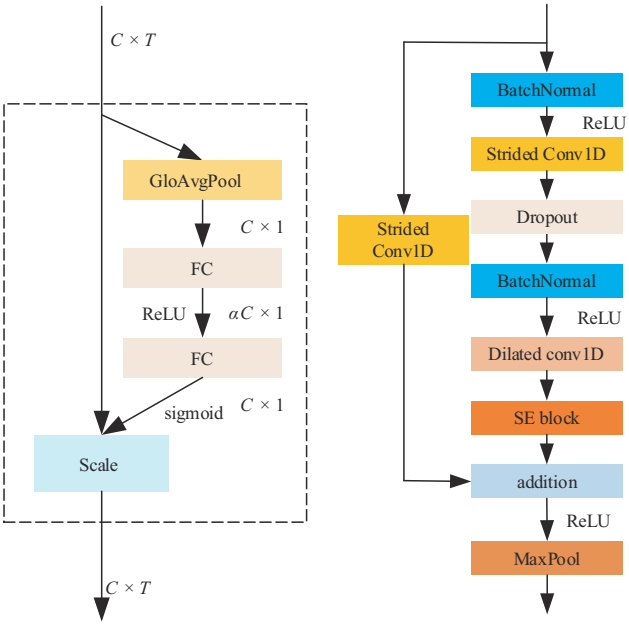
2.5　众数决策

对于一条待测试音频,按照第2.1节分割训练集的方法分割为多个片段,使用神经网络逐个对片段进行,将分类结果的众数作为该待测试语音的标签. 当分类结果中各类的数量相同时,将所有片段的分类结果按照对应的维度相加,而后分别在对应维度求平均,平均值最大的维度即为相应的类别.

表1　Res-SE sampleCNN模型各层输出

输入音频长度为6 s(大小为1×96 000)		
层名称	输出大小	参数设置
Strided Conv1D 3-64	64×32 000	stride=3,dilation=1
Dilated Conv1D 3-64	64×32 000	stride=1,dilation=2
Dilated Conv1D 3-64	64×32 000	stride=1,dilation=2
Res-SE block	128×8 000	
Res-SE block	256×2 000	
Res-SE block	512×500	
Res-SE block	1 024×125	
Res-SE block	2 048×15	
FC	2 048×1	
FC	2 048×1	
softmax	3	

注:"Conv1D 3-64"中,"3"表示滤波器的大小,"64"表示使用64个滤波器.



(a) SE模块 (b) Res-SE模块

图4 SE模块和Res-SE模块

3 实验设置与结果分析

3.1 数据集

我们使用NCMMSC2021 AD和INTERSPEECH-2020 ADReSS数据集分别对模型进行测试. 表2、表3分别是两个数据集的统计信息. NCMMSC2021 AD训练集中包含280条语音片段,分别为阿尔茨海默症患者的79个语音片段、轻度认知障碍患者的93个语音片段和健康人的108个语音片段. 每条数据时长约28~60 s,总时长约4.18 h,数据的采样率为16 kHz. 语音内容包括图片描述、流畅性测试、自我介绍等. 测试集中包含119个时长44~60 s的长音段和1 153个时长6 s的短音段,其中长、短音段测试集阿尔茨海默症患者、轻度认知障碍患者、健康人的数量均为10名、23名、20名. 训练集与测试集中的语音数据的语种均为汉语.

表2 NCMMSC2021 AD数据集统计信息

类别		音频数量	时长/s	总时长/h
训练集	AD	79	29~60	1.20
	MCI	93	28~60	1.62
	HC	108	28~60	1.36
测试集	长音频	119	44~60	1.94
	短音频	1 153	6	1.92

INTERSPEECH2020 ADReSS数据集共有156条语音数据,阿尔茨海默症和非阿尔茨海默症(non-AD)的语音数量都是78条,分别来自78名患者和78名健康老人,内容主要为图片描述和讲话录音等[25]. 每条数据时长约26~268 s,总时长约3.24 h. 训练集中阿尔茨海默症和非阿尔茨海默症的音频数量均为54条. 测试集包含24名AD患者和24名non-AD的语音数据. 语音数据的语种为英语.

表3 INTERSPEECH2020 ADReSS数据集统计信息

类别		音频数量	时长/s	总时长/h
训练集	AD	54	28~224	1.23
	non-AD	54	30~135	0.93
测试集		48	26~268	1.08

3.2 实验设置

由于训练数据长度不同,且NCMMSC2021 AD短音频测试集时长为6 s,因此我们使用长度为6 s、窗移为2 s的时间窗对训练数据进行分割,分割后音频片段的标签与原音频的标签相同,由此将NCMMSC2021 AD训练集的280条音频分割为6 929条训练数据. 对于长音频赛道的119条测试数据,使用相同的方法进行分割,分割后得到3 245条数据. ADReSS数据集分割后得到3 626条训练集数据和1 860条测试集数据. 在测试时,对分割后的音频片段分别进行测试,将测试结果的众数作为待测试音频的类别. 按照9∶1的比例将训练集分割为训练集、验证集,训练完成后在测试集上进行测试.

在输入为原始波形的实验中,我们使用Adam优化器更新神经网络的权重,学习率设置为0.000 1,批大小为32[7],并且在最后一个全连接层之前插入了一个参数为0.5的dropout层[26],挤压-激励模块的超参数α为1/16[21];而在输入为梅尔谱的实验中,使用帧长20 ms、帧移10 ms的汉宁窗对分割后的音频片段进行分帧之后[27],使用torchlibrosa工具提取64维的对数梅尔谱图,将特征送入网络中进行训练,优化器的参数和批大小与输入为原始波形的实验相同. 需要指出的是,实验中的Model 1仅使用了文献[28]中的CNN的网络结构,未使用预训练模型进行实验.

3.3 结果分析

为较为全面地评价模型的分类效果,我们以准确率作为评价指标,将采用Macro规则计算得到的精度、召回率、$F1$值作为补充指标.

从NCMMSC2021 AD长音频测试集的结果(表4)可以看出,以原始波形作为输入的模型,准确率大多高于Model 4基线系统的准确率,并且Model 8、Model 9的准确率要高过基于梅尔谱图的模型,这表明使用残差结构的sampleCNN模型可能从原始波形中提取到了比手工设计特征更丰富的信息. 在短音频测试结果(表5)中,Model 8和Model 9的准确率分别比基线高6.83%、7.35%,也高过以梅尔谱图作为输入的模型. 在输入都是原始波形的情况下,与一维的CNN、ResNet模型相比,

Res-SE sampleCNN也表现出了较好的分类性能. 整体而言,长音频测试集的准确率普遍高于短音频,我们推测可能是由于AD、MCI患者语音特征不稳定,当语音片段时长较短时,患者和正常人之间的语音特征容易产生混淆. 表6是各模型在ADReSS数据集上的结果. 从实验结果可以看出,Res-SE sampleCNN模型比基线提高4.17%,说明模型对于阿尔茨海默症和非阿尔茨海默症具有较好的区分度.

图5、图6分别是Model 9在NCMMSC2021 AD长音频、短音频测试集上的混淆矩阵,从图中可以看到,长音频测试集中,HC的召回率最高,说明系统能够很好区分健康人和AD、MCI患者,但是图5和图6中MCI召回率最低,并且模型很容易将MCI判断为AD,可能是由于AD和MCI的语音特征之间有较多重合. 图7(a)、图7(b)分别为Model 9在NCMMSC2021 AD和INTERSPEECH2020 ADReSS数据集上,训练集和验证集的损失函数变化情况.

表4　NCMMSC2021 AD长音频测试集实验结果

输入		模型	准确率	召回率	精度	$F1$值
梅尔谱	Model 1	CNN14[28]	0.798 3	0.783 6	0.793 9	0.782 2
	Model 2	ResNet[19]	0.815 1	0.799 4	0.804 5	0.798 1
	Model 3	MobileNetV3[29]	0.823 5	0.808 0	0.814 4	0.806 2
eGeMAPS[13]	Model 4	SVM(baseline)	0.798 0	0.785 0	0.799 0	0.786 0
原始波形	Model 5	1D-CNN	0.773 1	0.779 3	0.790 1	0.773 4
	Model 6	1D-ResNet	0.806 7	0.808 8	0.806 8	0.806 0
	Model 7	sampleCNN[17]	0.806 7	0.802 9	0.829 8	0.796 4
	Model 8	Res-sampleCNN	0.848 7	0.854 1	0.863 1	0.846 2
	Model 9	Res-SE sampleCNN	**0.865 5**	0.861 8	0.867 4	0.861 7

注:Model 4是NCMMSC2021 AD比赛的长音频测试集基线,链接为https://github.com/THUsatlab/AD2021;Model 5的地址为https://github.com/Jumabek/net_intrusion_detection,使用其中的一维CNN5模型进行实验;Model 6的地址为https://github.com/geekfeiw/Multi-Scale-1D-ResNet,使用其中卷积核大小为3的支路进行实验.

表5　NCMMSC2021 AD短音频测试集实验结果

输入		模型	准确率	召回率	精度	$F1$值
梅尔谱	Model 1	CNN14[28]	0.773 6	0.759 4	0.763 1	0.756 2
	Model 2	ResNet[19]	0.790 1	0.776 3	0.781 5	0.773 5
	Model 3	MobileNetV3[29]	0.779 7	0.763 9	0.771 5	0.759 2
梅尔谱	Model 4	CNN(baseline)	0.740 0	0.737 0	0.723 0	0.718 0
原始波形	Model 5	1D-CNN	0.749 3	0.751 5	0.758 0	0.748 5
	Model 6	1D-ResNet	0.784 0	0.781 4	0.781 5	0.781 4
	Model 7	sampleCNN[17]	0.780 6	0.776 7	0.801 1	0.773 2
	Model 8	Res-sampleCNN	0.808 3	0.808 8	0.811 8	0.805 9
	Model 9	Res-SE sampleCNN	**0.813 5**	0.814 2	0.814 2	0.811 9

注:Model 4是NCMMSC2021 AD比赛的短音频测试集基线,链接为https://github.com/THUsatlab/AD2021.

表6　INTERSPEECH2020 ADReSS数据集实验结果

输入		模型	准确率	召回率	精度	$F1$值
梅尔谱	Model 1	CNN14[27]	0.583 3	0.583 3	0.588 9	0.576 7
	Model 2	ResNet[19]	0.625 0	0.625 0	0.625 9	0.624 3
	Model 3	MobileNetV3[29]	0.645 8	0.645 8	0.659 3	0.638 1
ComParE[30]	Model 4	LDA(baseline)[25]	0.625 0	0.625 0	0.635 0	0.620 0
原始波形	Model 5	1D-CNN	0.604 2	0.604 2	0.604 3	0.604 0
	Model 6	1D-ResNet	0.562 5	0.562 5	0.563 5	0.560 8
	Model 7	sampleCNN[17]	0.583 3	0.583 3	0.585 7	0.580 4
	Model 8	Res-sampleCNN	0.645 8	0.645 8	0.646 1	0.645 7
	Model 9	Res-SE sampleCNN	**0.666 7**	0.666 7	0.701 7	0.651 5

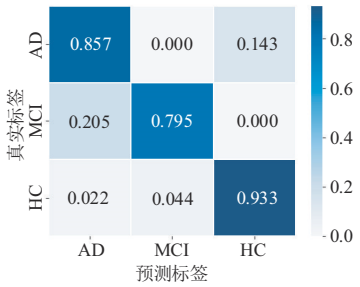


图5 Model 9在NCMMSC2021 AD长音频测试集的混淆矩阵

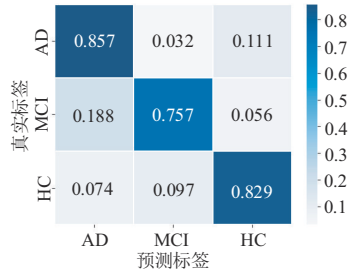
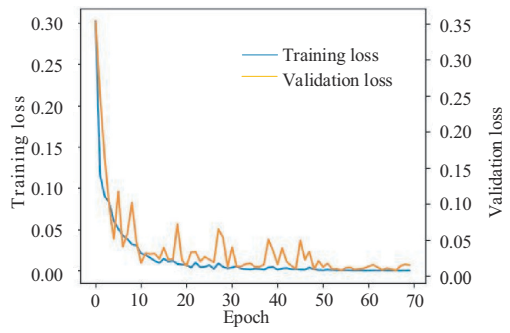
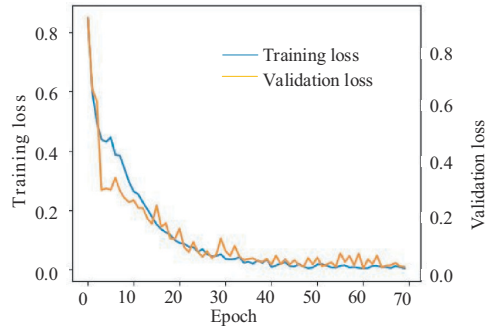


图6 Model 9在NCMMSC2021 AD短音频测试集的混淆矩阵



(*a*) NCMMSC2021 AD



(*b*) ADReSS

图7 Model 9在数据集上的损失函数变化

通常在使用CNN、ResNet等神经网络处理语音数据时,先将语音数据转换为二维的频域表示,因此会丢失部分时间维度的信息,而一维卷积可以好的处理时序数据,提取时间维度的特征[31]. 对于阿尔茨海默症患者来讲,其语音的时序特征较为显著,话语间的停顿、重复、中断时长等与健康人不同,因此研究者使用停顿间隔、词的重复频率等作为特征进行分类[9,10]. 在Model 8中,通过使用含有膨胀卷积的残差块增加模型的深度和感受野,从而提取了较大时间尺度的特征. 而在Model 9中,加入的挤压-激励模块促使模型学习不同通道的信息,使得准确率进一步提高.

相比于二维卷积,当网络层数相同时,一维卷积的参数量更少,并且只需要计算标量的乘法和加法,因此计算的复杂度较低,需要的资源较少[31]. CNN、ResNet等神经网络参数较多,需要大量的训练数据才能获得较好的性能. 对于阿尔茨海默症检测来说,由于医学伦理等制约,难以获得大量的训练数据,可能使用一维卷积是较好的选择.

上述分析表明,对于sampleCNN模型而言,将卷积-池化模块替换为含有膨胀卷积的残差块,并在残差块中加入挤压-激励模块后,准确率有明显提高. 最终模型与基线相比,在NCMMSC2021 AD的长音频测试集、短音频测试集上分别提升6.75%、7.35%,在INTERSPEECH2020 ADReSS数据集上提升4.17%,通过实验证实,基于原始波形的Res-SE sampleCNN方法有助于实现更好的阿尔茨海默症检测.

4 结论

本文提出了一种基于原始波形的端到端阿尔茨海默症检测方法,在NCMMSC2021 AD和INTERSPEECH2020 ADReSS数据集上进行了测试. 我们对实验数据进行了数据增强,将sampleCNN中使用的卷积块替换为含有膨胀卷积的残差块,同时在残差块中加入挤压-激励模块以提高模型的分类能力,实验证明我们提出的模型对于阿尔茨海默症患者具有较好的分类性能. 未来,我们将探索不同语种的阿尔茨海默症患者语音特征的差异,并研究不同尺度的特征对于模型性能的影响.

参考文献

[1] MATTSON M P. Pathways towards and away from Alzheimer's disease[J]. Nature, 2004, 430(7000): 631-639.

[2] WORLD HEALTH ORGANIZATION. Dementia: A public helath priority[EB/OL]. (2012-04-11)[2022-01-06]. https://www.who.int/mental_health/publications/dementia_report_2012/en/.

[3] FOLSTEIN M F, FOLSTEIN S E, MCHUGH P R. "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician [J]. Journal of Psychiatric Research, 1975, 12(3): 189-198.

[4] NASREDDINE Z S, PHILLIPS N A, BÉDIRIAN V, et al. The Montreal cognitive assessment, MoCA: A brief screen-

ing tool for mild cognitive impairment[J]. Journal of the American Geriatrics Society, 2005, 53(4): 695-699.

[5] SELKOE D J. Alzheimer's disease[J]. Cold Spring Harbor Perspectives in Biology, 2011, 3(7): a004457.

[6] MUELLER K D, KOSCIK R L, HERMANN B P, et al. Declines in connected language are associated with very early mild cognitive impairment: Results from the Wisconsin registry for Alzheimer's prevention[J]. Frontiers in Aging Neuroscience, 2018, 9: 437.

[7] CHIEN Y W, HONG S Y, CHEAH W T, et al. An automatic assessment system for Alzheimer's disease based on speech using feature sequence generator and recurrent neural network[J]. Scientific Reports, 2019, 9(1): 19597.

[8] CHEN J, YE J P, TANG F Y, et al. Automatic detection of Alzheimer's disease using spontaneous speech only[C]//Interspeech, 2021. Baixas: International Speech Communication Association, 2021: 3830-3834.

[9] KÖNIG A, SATT A, SORIN A, et al. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease[J]. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 2015, 1(1): 112-124.

[10] LUZ S. Longitudinal monitoring and detection of Alzheimer's type dementia from spontaneous speech data [C]//2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS). Piscataway: IEEE, 2017: 45-46.

[11] YUAN J H, BIAN Y C, CAI X Y, et al. Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease[C]//Interspeech 2020. Baixas: International Speech Communication Association, 2020: 2162-2166.

[12] ZHU Z N, NOVIKOVA J, RUDZICZ F. Detecting cognitive impairments by agreeing on interpretations of linguistic features[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. Stroudsburg: Association for Computational Linguistics, 2019: 1431-1441.

[13] EYBEN F, SCHERER K R, SCHULLER B W, et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing[J]. IEEE Transactions on Affective Computing, 2015, 7(2): 190-202.

[14] PALAZ D, COLLOBERT R, MAGIMAI-DOSS M. Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks[C]//Interspeech 2013. Baixas: International Speech Communication Association, 2013: 1766-1770.

[15] MUCKENHIRN H, MAGIMAI-DOSS M, MARCELL S. Towards directly modeling raw speech signal for speaker verification using CNN S[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2018: 4884-4888.

[16] HOSHEN Y, WEISS R J, WILSON K W. Speech acoustic modeling from raw multichannel waveforms[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2015: 4624-4628.

[17] LEE J, PARK J, KIM K L, et al. Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms[EB/OL]. (2017-03-06) [2022-01-06]. https://arxiv.org/abs/1703.01789.

[18] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04) [2022-01-06]. https://arxiv.org/abs/1409.1556.

[19] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.

[20] YU F, KOLTUN V, FUNKHOUSER T. Dilated residual networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 636-644.

[21] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.

[22] GRAVES A, MOHAMED A R, HINTON G. Speech recognition with deep recurrent neural networks[C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2013: 6645-6649.

[23] HE K M, ZHANG X Y, REN S Q, et al. Identity mappings in deep residual networks[C]//14th European Conference on Computer Vision (ECCV). Berlin: Springer Verlag, 2016: 630-645.

[24] KIM T, LEE J, NAM J. Comparison and analysis of SampleCNN architectures for audio classification[J]. IEEE Journal of Selected Topics in Signal Processing, 2019, 13 (2): 285-297.

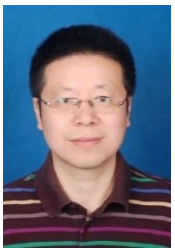
[25] LUZ S, HAIDER F, DE LA FUENTE S, et al. Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge[C]//Interspeech 2020. Baixas: In-

ternational Speech Communication Association, 2020: 2172-2176.
[26] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014,15(1):1929-1958.
[27] PRIYANKA M A S, SOLOMI V S, VIJAYALAKSHMI P, et al. Multiresolution feature extraction (MRFE) based speech recognition system[C]//2013 International Conference on Recent Trends in Information Technology (ICRTIT). Piscataway: IEEE, 2014: 152-156.
[28] KONG Q Q, CAO Y, IQBAL T, et al. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2880-2894.
[29] HOWARD A, SANDLER M, CHEN B, et al. Searching for MobileNetV3[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2020: 1314-1324.
[30] EYBEN F, WENINGER F, GROSS F, et al. Recent developments in openSMILE, the Munich open-source multimedia feature extractor[C]//Proceedings of the 21st ACM international conference on Multimedia. New York: ACM, 2013: 835-838.
[31] KIRANYAZ S, AVCI O, ABDELJABER O, et al. 1D convolutional neural networks and applications: A survey [J]. Mechanical Systems and Signal Processing, 2021, 151: 107398.

作者简介

陈旭初 男,1992年12月出生于河南省驻马店市. 现为清华大学电子工程系在读硕士研究生. 主要研究方向为音频事件检测、情感识别.
E-mail: chen-xc20@mails.tsinghua.edu.cn

张卫强(通讯作者) 男,1979年1月出生于河北省雄县. 2002年于中国石油大学应用物理系获学士学位,2005年于北京理工大学电子工程系获硕士学位,2009年于清华大学电子工程系获博士学位,2017年斯坦福大学访问学者,现为清华大学电子工程系副研究员. 主要研究方向为语音与音频信号处理.
E-mail: wqzhang@tsinghua.edu.cn



马 勇 男,1977年3月出生于江苏省新沂市. 2015年于北京工业大学电子信息与控制工程学院获得博士学位. 现为江苏师范大学讲师,从事语音与音频信号处理方面的研究工作.
E-mail: may@jsnu.edu.cn