王茂, 何勇. 基于 FBank 特征与改进 CNN 的声纹识别研究[J]. 智能计算机与应用,2024,14(8):40-47. DOI:10.20169/j.issn.2095-2163.240807

基于 FBank 特征与改进 CNN 的声纹识别研究

王 茂, 何 勇
(贵州大学 计算机科学与技术学院, 贵阳 550025)

摘 要: 在声纹识别研究中,针对声纹信号特征的表征能力不足,模型的识别准确率不高的问题,提出基于 CNN 卷积神经网络的声纹识别方法,使用能体现更多声音本质的 FBank 梅尔语谱图特征作为模型的输入;此外,大多数研究为提高识别率而广泛使用多层堆叠的网络结构,使得网络参数量与计算量较大,难以部署到计算资源和存储资源紧缺的边缘智能设备。为此,提出采用分组卷积的方式对 CNN 标准主干网络结构进行优化,降低网络参数量和计算量;同时为了保证网络模型的识别准确率,采用 CBAM 注意力机制进一步优化网络,使其关注通道和空间中更有价值的地方。经实验验证,所提方法有较高的声纹识别准确率,且相较于标准 CNN,优化后的模型参数量与计算量均有较大程度的减少。

关键词: 声纹识别; 卷积神经网络; FBank 特征; 分组卷积; CBAM 注意力机制
中图分类号: TP391.4 **文献标志码**: A **文章编号**: 2095-2163(2024)08-0040-08

Research on voiceprint recognition based on FBank feature and improved CNN

WANG Mao, HE Yong
(College of Computer Science and Technology, Guizhou University, Guiyang 550025, China)

Abstract: In the study of voice print recognition, on the one hand, aiming at the problems of insufficient characterization ability of voice print signal features and low recognition accuracy of the model, a voice print recognition method based on CNN convolutional neural network is proposed, and FBank Meir spectrogram features that can reflect more sound essence are used as the input of the model. On the other hand, in order to improve the recognition rate, the multi-layer stacked network structure is widely used in most current researches, which makes the number of network parameters and FLOPS(floating point of per second) large, and difficult to deploy to the edge intelligent devices with scarce computing resources and storage resources. Therefore, a grouping convolution method is proposed to optimize the CNN standard backbone network structure to reduce the number of network parameters and the amount of FLOPS. At the same time, in order to ensure the recognition accuracy of the network model, CBAM attention mechanism is used to further optimize the network and make it focus on more valuable places in the channel and space. The experimental results show that the proposed method has a higher voire recognition accuracy, and compared with standard CNN, the number of parameters and FLOPS of the optimized model are greatly reduced.

Key words: voice print recognition; convolutional neural network; FBank characteristics; grouping convolution; CBAM attention mechanism

0 引 言

声纹识别作为一种生物体个体识别方式,可以传达一个人的特征相关的信息,识别说话人的身份[1]。与其他生物特征识别技术相比,声纹识别具有非接触、操作简单、成本节约等优点,因而被普遍运用于各个领域。例如:公安部门刑事侦查时,采用声纹识别进行说话人辨认,以便缩小犯罪嫌疑人的搜查范围;银行认证或者手机语音助手识别机主时,使用声纹识别来对目标人物的身份进行确认等。

由于声纹识别技术被广泛运用于各个领域,因此其发展也引起了人们的高度重视[2]。在声纹识别领域中,传统的主流识别模型有:高斯混合-通用背景模型 (Gaussian Mixture Model - Universal

基金项目: 贵州省科技计划项目(黔科合支撑[2020]2Y007 号)。
作者简介: 王 茂(1998-),女,硕士研究生,主要研究方向:声纹识别,深度学习。
通讯作者: 何 勇(1974-),男,博士,副教授,硕士生导师,主要研究方向:物联网,嵌入式系统。Email:2055763134@qq.com
收稿日期: 2023-05-22

Background Model，GMM-UBM）[3]、高斯混合-支持向量机模型（Gaussian Mixture Model-Support Vector Machine，GMM-SVM）[4]、联合因子分析（Joint Factor Analysis，JFA）[5]、i-vecto[6]等。这类声纹识别模型提取和使用的主要声纹特征是梅尔频率倒谱系数 MFCC[7]。该特征在提取过程中经过了 DTC 线性变换，损失了声音信号中原有的部分非线性特征，导致表征能力不够强；且此类模型忽略了说话人特征与信道特征的相关性，缺乏对信道多变性的补偿能力，导致泛化能力不强，鲁棒性差，识别准确率不高[8]。

随着深度学习技术的不断发展，体现出了各种优良特性。如：可以进行深层次的非线性变换，提取足够多的特征，使模型具有超强的特征学习和表达能力；能够通过分层特性提升特征的鲁棒性，使模型不容易受到噪声信号的干扰；易迁移性使模型具有较强的泛化能力等。这些特性使得深度学习技术非常适合进行复杂声音信号的处理和建模，因此学术界存在较多的基于 LSTM、RNN 以及 CNN 等深度学习网络模型进行声纹分类识别的研究：如文献[9]根据深度学习在语音识别中的优缺点，分析了语音识别技术的发展前景和挑战；文献[10]通过多层 TDNN（时间延迟神经网络）和 LSTM（长短期记忆）深度神经网络的组合实现语音识别；文献[11]提取 Mel-Frequency 倒谱系数进行评估，使用支持向量机在两个不同的数据集上进行训练和测试等。但是这些深度学习神经网络模型在进行分类识别时，为了取得更好的识别效果，一般会不断增加网络模型卷积层的深度与宽度，使得模型参数量与计算量过大，难以在一些存储和计算资源受限制的边缘智能设备上部署应用。

针对以上问题，本文提出采用 FBank 梅尔语谱图作为表征声音信号的特征，搭建经过改进和优化的卷积神经网络模型（Convolutional Neural Network，CNN）作为说话人识别模型。即：在标准 CNN 结构基础上，通过分组卷积的方式优化 CNN 标准卷积结构。同时采用注意力机制[12]（Convolutional Block Attention Module，CBAM）对网络进行进一步优化，使得网络关注更加重要的特征，在达到较高识别效率的同时减少参数量、计算量以及网络模型大小。

1　声纹特征与卷积神经网络

1.1　声纹特征

根据奈奎斯特定理，在进行采样时，采样频率需要达到 2 倍以上的最高信号采集频率，才能确保信号不会失真[13]。本文使用的语音信号数据集为保证信号质量，将采样频率设置为 16 kHz，即每秒约 16 000 byte的样本数据量。数据量较大，若直接输入神经网络模型会增大网络模型的负担，因此通常先提取语音数据的声纹特征，然后将其作为模型的输入或输出。

本文提取语音数据的 FBank 梅尔语谱图作为声纹特征。FBank 特征提取算法类似于人耳的方式对音频进行处理，拟合人耳接收声音信号的特性，因此提取到的声纹特征更多的保留了声音信号的本质。与传统模型常使用的 MFCC 特征相比，FBank 特征没有应用离散余弦变化进行去相关处理，其计算量更小且特征相关度更高，包含更多的信息。将其用于声纹识别，能够达到更好的识别准确率。FBank 特征提取算法过程如图 1 所示：首先进行预加重对语音高频部分进行增益，然后进行分帧及加窗得到短时信号，最后再经过短时傅里叶变换（STFT）、Mel 滤波、取对数等运算得到最终的 FBank 梅尔语谱图特征。

图 1　FBank 特征提取流程

Fig.1　FBank feature extraction process

由于上述得到的信号仅仅是语音在时域上的表现，而看不出其在频域上的特征，因此接下来通过 STFT 将信号对应地转换为频域上的能量分布。具体实现过程如下：

（1）短时傅里叶变换（STFT）过程如式（1）所示：

$$\mathrm{STFT}(t,f)=\int_{-¥}^{¥}x(\tau)h(\tau-t)e^{-\mathrm{j}2\pi f\tau}\mathrm{d}\tau \quad (1)$$

（2）计算能量谱

经过 STFT 变换后的信号是频域信号，由于其在不同的频带范围内的能量大小不同，因此需要分别计算出语音信号对应的能量谱。

（3）Mel 滤波

由于人耳对低频的语音信号比较敏感，对高频信号相对不敏感。为了模拟人耳的这种特性，将频率映射到梅尔频率，梅尔频率与频率之间的关系如式（2）所示：

$$\mathrm{Mel}(f)=2\,595\cdot\ln\left(1+\frac{f}{700}\right) \quad (2)$$

式中：f 代表原本的频率，Mel 代表转换后的梅尔频率。此外，为了使得频谱更加平滑，同时消除谐波对信号的影响，接下来对信号进行 Mel 滤波：首先定义一组滤波器组，然后将步骤(3)生成的能量谱通过该滤波器组进行滤波。三角滤波器的频率响应定义为：

$$H_m(k)=\begin{cases}0, & k<f(m-1)\\ \dfrac{2(k-f(m-1))}{(f(m+1)-f(m-1))(f(m)-f(m-1))}, & f(m-1)\leqslant k\leqslant f(m)\\ \dfrac{2(f(m+1)-k)}{(f(m+1)-f(m-1))(f(m)-f(m-1))}, & f(m)\leqslant k\leqslant f(m+1)\\ 0, & k\geqslant f(m+1)\end{cases} \tag{3}$$

式中：$\sum_{m=0}^{M-1}H_m(k)=1$。

(4)对数运算

得到 Mel 滤波结果后，对其进行取对数操作，将信号在低能量处的差异放大。取对数的计算公式如下：

$$s(m)=\ln\left(\sum_{k=0}^{N-1}|X_a(k)|^2H_m(k)\right),\ 0\leqslant m\leqslant M \tag{4}$$

经过对数运算后得到的结果即为 FBank 梅尔语谱图特征，最终将其以图像格式作为神经网络模型的特征输入。

1.2 卷积神经网络

CNN 卷积神经网络是一种经典的深度学习模型，其主要用于图像识别、图像分类、目标检测等计算机视觉领域[14]。其在计算机视觉中的成功应用使其成为了当前最为流行的深度学习模型之一。

CNN 的核心思想是利用卷积操作进行特征提取。卷积操作是一种局部感知机制，其可以通过滑动一个卷积核在输入图像上进行卷积计算，从而提取图像中的局部特征。卷积操作的输出称为特征图(Feature Map)，并且通过不同的卷积核，可以提取出不同的特征。CNN 通常包含多个卷积层、池化层、全连接层等模块。其中卷积层用于提取特征，池化层用于降低特征图的空间尺寸，全连接层用于分类或回归任务。其主干网络结构如图 2 所示。

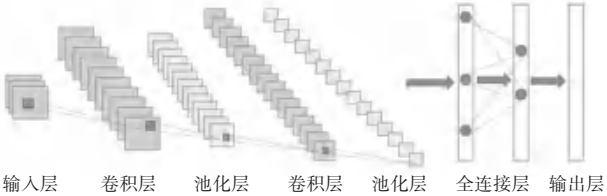


图 2 CNN 网络结构图

Fig. 2 CNN network structure diagram

在卷积层中，卷积核会对输入数据进行卷积计算，然后通过激活函数进行非线性变换。在池化层中，常见的方法包括最大池化和平均池化，其可以有效地减少特征图的空间尺寸。在全连接层中，所有特征都会被压缩成一个向量，并通过 Softmax 函数进行分类。CNN 在图像识别、目标检测等领域表现优异，其主要优势在于其可以利用卷积操作进行局部感知和参数共享，从而提高模型的准确性和泛化能力，因此本文也以 CNN 为基础，建立识别模型对人的语音进行识别。

2 改进的声纹识别算法

针对 CNN 标准网络结构应用于声纹识别领域时存在参数量与计算量大、准确率不高的情况，本文对网络进行了以下优化：一是在不损失精度的前提下，对网络结构进行分组优化以减小参数量、计算量；二是通过 CBAM 注意力机制，在几乎不增加网络负担的情况下，使网络关注通道和空间中更需要关注的地方，进一步增强 FBank 特征的表征能力，增加说话人身份识别的精确率。

2.1 分组卷积优化网络结构

CNN 标准卷积是一种全通道卷积，即每个卷积核分别与所有的输入特征图组卷积，这样的卷积方式产生的参数较多。因此本文采用分组卷积[15]的方式，对标准主干网络卷积层进行优化，分组卷积结构如图 3 所示。将输入层的特征图分为 G 组，对应 G 个过滤器组，每个过滤器组包含 D_{out}/G 个过滤器，即最终每个过滤器组都将输出 D_{out}/G 个通道，整体共输出的通道数同标准卷积一样为：$G*D_{out}/G=D_{out}$。但在每个过滤器组中，其深度仅为标准卷积的 D_{in}/G，大大降低了卷积的计算量和参数量。

两种卷积方式具体的参数量与计算量计算过程如下。标准卷积的参数量计算过程如式(5)所示，标准卷积的计算量计算如式(6)所示。

$$P_{SC}=K^2\times C_0\times C_1 \tag{5}$$

$$F_{SC}=(2\times K^2\times C_0-1)\times H\times W\times C_1 \tag{6}$$

其中，K 为卷积核大小；C_0 为输入通道数；C_1 为输出通道数；H、W 表示输出特征图的尺寸大小。

采用分组卷积对输入特征图进行卷积时的参数

量计算过程如式(7)所示，分组卷积计算量的计算如式(8)所示。

$$P_{GC} = K^2 \times \frac{C_0}{G} \times \frac{C_1}{G} \times G \tag{7}$$

$$F_{GC} = \left(2 \times K^2 \times \frac{C_0}{G} + 1\right) \times H \times W \times \frac{C_1}{G} \times G \tag{8}$$

其中各符号含义与公式(5)、公式(6)相同。

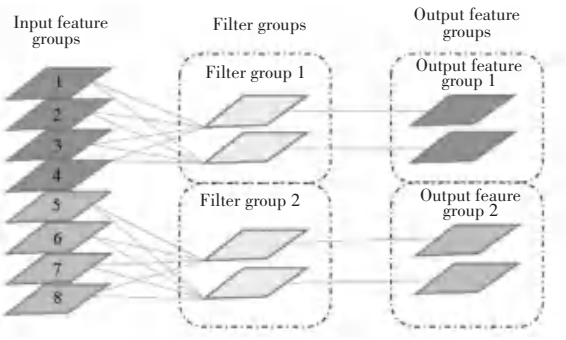


图 3　分组卷积结构示意图

Fig. 3　Schematic diagram of grouping convolution structure

根据以上公式可以推出：经过分组卷积后网络的计算量和参数量与分组数 G 成反比，G 越大，计算量与参数量越小。因此本文在标准 CNN 网络结构的基础上构建了改进的 CNN 网络结构。

首先搭建基础的标准网络结构，其中包括 3 个卷积模块、一个全连接模块，每一个卷积模块后紧跟一个 2×2 的 MaxPooling 池化层以减少参数量。网络各层参数设置为：Convolution1 采用 24 个 5×5 的卷积核，步长 stride 为 1，填充尺寸 padding 为 2；Convolution2 与 Convolution3 分别采用 64、128 个大小为 5×5 的卷积核，stride 及 padding 与 Convolution1 一致。损失函数采用交叉熵函数 CrossEntropyLoss。在这个基础网络结构的基础上，保持卷积核个数、池化层大小、卷积步长等参数不变，对 3 个卷积模块的卷积层进行分组卷积。Convolution1 分为 3 组，Convolution2 分为 8 组，Convolution3 分为 8 组。经过分组后，每个卷积层的参数量与计算量均为标准卷积的 $1/G$。

2.2　CBAM 提高识别准确率

采用分组卷积的方式优化的 CNN 网络，由于减少了网络的参数量和计算量，使得每个组内的卷积核参数只能共享在组内，而不能跨组共享，这可能会导致网络的表达能力受限。

因此，本文采用 CBAM 注意力机制对分组卷积造成的这一缺陷进行补足。CBAM 是一种通用的注意力机制，适用于卷积神经网络（CNN）中的不同层[12]。CBAM 的主要目的是根据不同的特征通道(Channel)和空间位置(Spatial)分别进行注意力加权，以提高 CNN 的性能和准确性。

如图 4 所示，CBAM 注意力机制由两个子模块组成：通道注意力模块和空间注意力模块。这两个子模块都是在卷积神经网络中的每个卷积层后添加的，其作用如下：

通道注意力模块的作用是让网络关注哪些通道对于特定任务是最重要的。其通过全局池化层将每个通道的特征图压缩成一个标量，并通过一个全连接层(fully connected layer)将其映射到一个较小的维度。这个映射向量可以视为每个通道的重要性分数，其经过 Softmax 函数进行归一化，以产生一个加权的特征图，然后与输入特征图相乘，以产生最终的特征图。

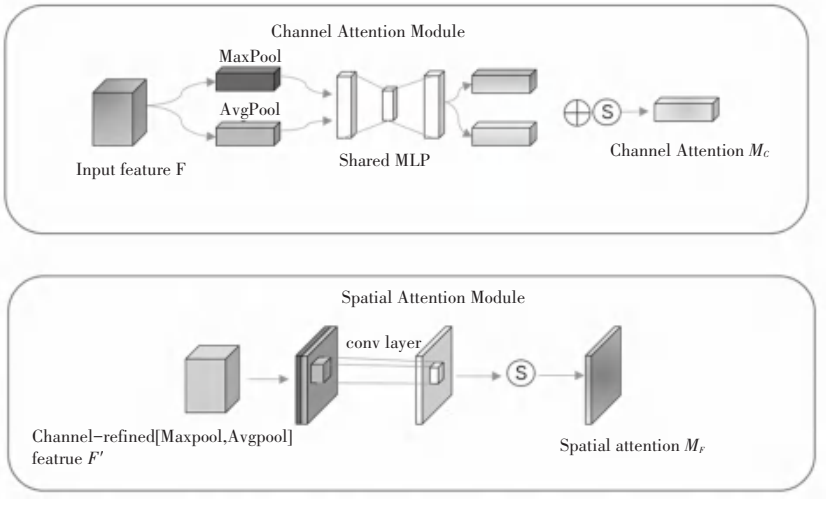


图 4　CBAM 算法示意图

Fig. 4　CBAM algorithm diagram

空间注意力模块的作用是让网络关注哪些空间位置对于特定任务是最重要的。其通过卷积层和池化层来捕捉特征图中的空间关系,并将其映射到一个加权的特征图。该模块包含两个分支:一个分支执行平均池化操作,另一个分支执行最大池化操作。这两个分支提取的特征图被级联起来,并通过卷积层进行处理,然后经过 Sigmoid 函数进行归一化,以产生一个空间注意力图。这个空间注意力图与输入特征图相乘,以产生最终的特征图。

最后,通道注意力模块和空间注意力模块的输出都被乘以归一化的缩放因子(Scaling Factor),以确保输入特征图的值范围不会发生显著的变化。缩放因子的计算方式:是将注意力加权特征图进行全局平均池化,然后将结果通过一个全连接层和一个 ReLU 函数进行处理,得到一个正的缩放因子。

总之,CBAM 注意力机制通过自适应地调整每个特征通道和空间位置的权重,可以增强卷积神经网络的特征表示能力。这种机制不会显著增加计算成本,因为其只是在每个卷积层后添加一个小的注意力模块。因此根据以上理论依据,本文在分组卷积网络的最后一个卷积模块加上了 CBAM 注意力机制,得到最终改进的 CNN 网络结构如图 5 所示。该网络能在几乎不增加参数量的情况下,使得网络的准确率得到进一步的提高。

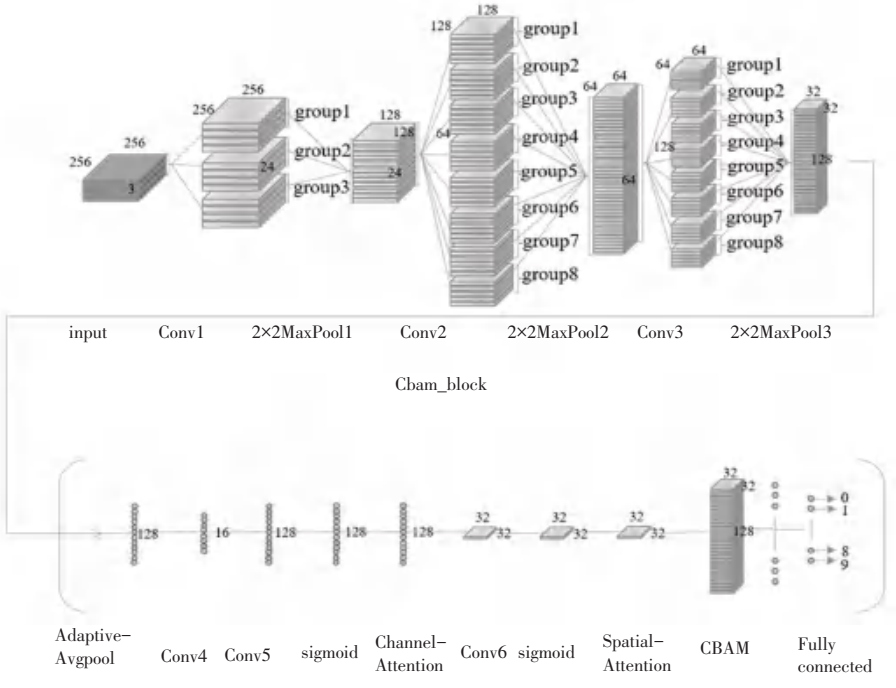


图 5 改进的 CNN 网络结构示意图

Fig. 5 Improved CNN network structure

3 实验验证及分析

3.1 实验平台

本实验在深度学习框架 Pytorch 下进行。操作系统为 Ubuntu16.04LTS。计算机硬件配置参数为:2080TiGPU、32 GB RAM、16 GB 显存。

3.2 实验数据集

实验采集的语音数据集为 Free ST Chinese Mandarin Corpus,由 Surfingtech 中文普通话语料库免费提供[16]。该数据集由 855 位说话人的语音组成,每位说话人包含 120 个时长为 3~4s 的语音片段。语音采样频率为 16 000 Hz。

本文实验分别将语料库中每位说话人的 120 个语音片段按照 8∶2 的比例划分出训练集和测试集,其中每位说话人的训练集语音片段个数为 96,测试集为 24。

3.3 实验设计

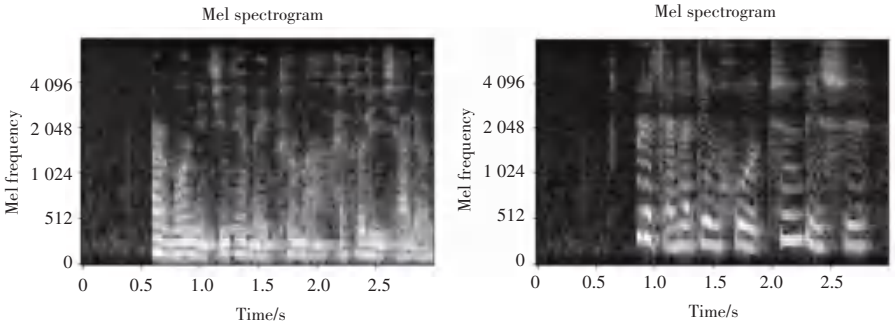
1)提取特征

如图 6 所示是不同说话人的语音提取出的 FBank 梅尔语谱图特征。由于不同说话人语音信号的梅尔语谱图存在差异,经过处理后输入网络模型中的每幅梅尔语谱图大小为 256×256。

2)对比实验

为了验证本文提出的声纹识别方法在减少网络

模型参数量、计算量的同时取得较高准确率的有效性，本文设计了 3 个对比实验：实验一采用标准的 CNN 网络，3 个卷积模块加上对应的 MaxPooling 池化层，最后是全连接层，提取及采用的特征为大量研究中常使用的 MFCC 特征；实验二同样采用标准的基础 CNN 网络结构，提取及采用的特征为 FBank 特征；实验三为改进后的 CNN 网络结构，提取及采用的特征为 FBank。

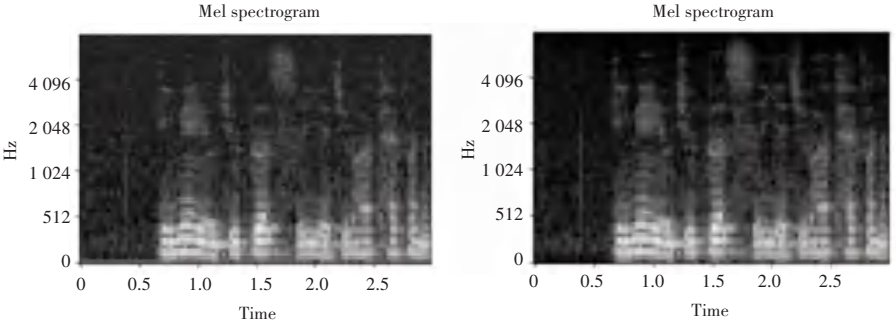


（a）说话人 A 梅尔语谱图　　（b）说话人 B 梅尔语谱图

图 6　两位说话人梅尔语谱图对比

Fig. 6　Comparison of Meir spectrogram between two speakers

为了进一步证明本文改进的 CNN 网络模型对声纹识别的有效性和鲁棒性，设计了实验四。该实验将前 3 个实验使用的测试集语音数据叠加高斯白噪声后，再提取相对应 FBank 特征，作为网络模型的测试集输入，加噪前后的 FBank 特征对比如图 7 所示。



（a）加噪前 FBank 特征　　（b）加噪后 FBank 特征

图 7　加噪前后 FBank 特征对比图

Fig. 7　Comparison of FBank features before and after noise

3.4　实验结果及分析

3.4.1　参数量与计算量分析

根据公式(5)~公式(8)以及实验得出改进后的 CNN 网络与标准 CNN 网络参数量、计算量及模型大小的对比见表 1，改进后的相对减少百分比见表 2。实验结果表明，优化后的 CNN 网络在一定程度上减少了计算资源与存储资源花销。

表 1　标准 CNN 与改进 CNN 各参数量

Table 1　Standard CNN and improved CNN parameters

模型	参数量	计算量	模型大小/MB
标准 CNN	1 555 946.0	12 698 255 360.0	5.952
改进 CNN	1 346 140.0	1 794 982 912.0	5.135

表 2　改进 CNN 计算量及参数量相对减少百分比

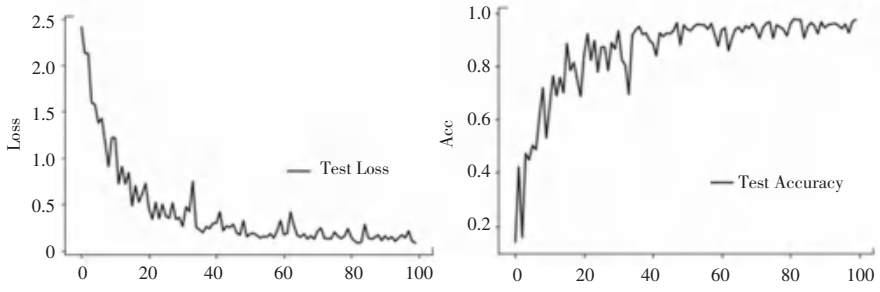
Table 2　Relative reduction percentage of the improved CNN FLOPS and params

对比参数	相对减少百分比/%
计算量	85.86
参数量	13.48
模型大小	13.73

3.4.2　损失值与准确率分析

对 3.2 小节划分的测试与训练数据集进行 MFCC、FBank 特征提取，然后分别使用实验一(CNN_MFCC)、实验二(CNN_FBank)、实验三(改进 CNN_FBank)中的网络模型进行训练，各模型的损失值以及准确率与迭代次数之间的关系如图 8~图 10 所示。通过对比 3 个实验的 Loss 曲线可以看到，3 组

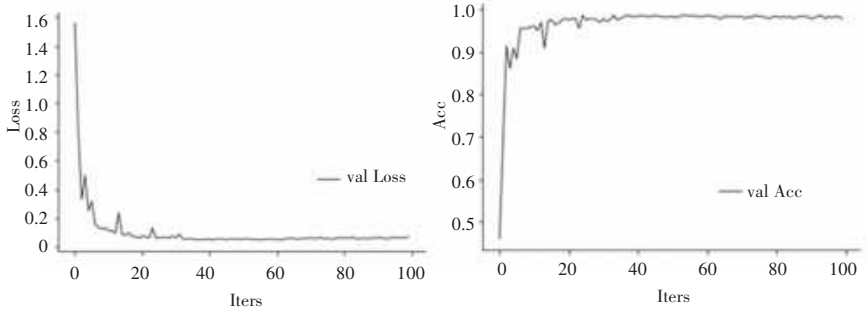
实验最终都能收敛,但实验三与实验二的收敛速度均比实验一的收敛速度更快;而实验二与实验三的收敛速度相当,说明 FBank 特征有更好的表征能力,使得模型得以快速收敛。



(a) 损失值变化曲线 (b) 准确率变化曲线

图 8 实验一损失值与准确率变化曲线

Fig. 8 Experiment 1 Loss value and Accuracy curve

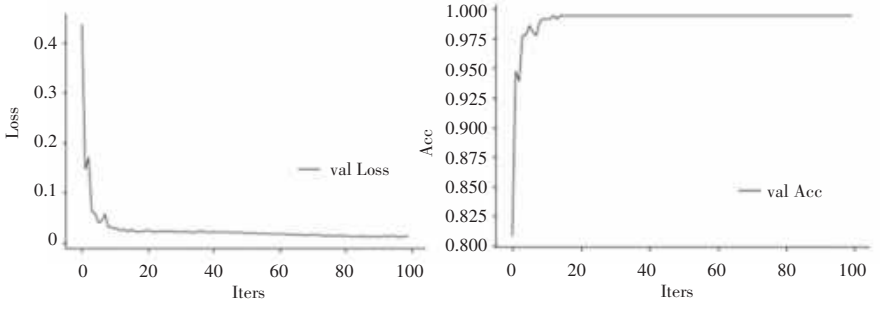


(a) 损失值变化曲线 (b) 准确率变化曲线

图 9 实验二损失值与准确率变化曲线

Fig. 9 Experiment 2 loss value and accuracy curve

通过对比 3 个实验的 *accuracy* 曲线(如图 8~图 10 所示)和最终的识别准确率(见表 3)可以看出,实验三的准确率比实验一高 2.66%,而实验三在参数量、计算量少于实验二的情况下,准确率比实验二高 0.72%,说明本文对标准网络进行分组卷积以及添加 CBAM 注意力机制的改进效果良好。



(a) 损失值变化曲线 (b) 准确率变化曲线

图 10 实验三损失值与准确率变化曲线

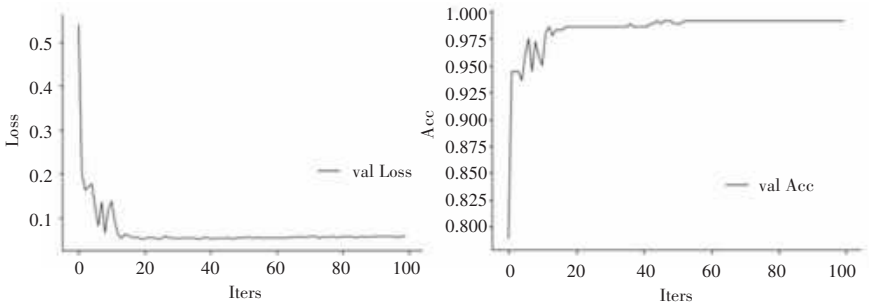
Fig. 10 Experiment 3 Loss value and Accuracy curve

表 3 各实验模型分类识别准确率

Table 3 Classification and recognition accuracy of each experimental model

Model	Acc/%
CNN-MFCC	96.41
CNN-FBank	98.35
改进 CNN-FBank	99.07
改进 CNN-加噪语音	97.22

同时,通过观察实验四测试集的准确率与损失值曲线(如图 11)可以看到,模型最终也能够逐渐收敛,并且准确率能够达到 97.22%(见表 3),说明本文提出的模型对声纹识别具有一定的有效性及鲁棒性。



（a）损失值变化曲线　　（b）准确率变化曲线

图 11　实验四损失值与准确率变化曲线

Fig. 11　Experiment 4 Loss value and Accuracy curve

4　结束语

本文将 CNN 应用于 FBank 梅尔语谱图实现声纹识别，从提高网络运算效率和增强声纹特征的表征能力 2 个方面对 CNN 进行改进，提出采用分组卷积以及增加 CBAN 注意力机制的方式实现对 CNN 的优化。实验结果表明，与优化前比较网络参数量减少了 13.48%，计算量减少了 85.86%，网络模型大小减少了 13.73%，准确率达到了 99.07%，且网络模型具有一定的抗干扰能力，在声纹识别方面具有较好的识别效果。

参考文献

[1] HANIFA R M, ISA K, MOHAMAD S. A review on speaker recognition: Technology and challenges[J]. Computers & Electrical Engineering, 2021, 90: 107005.

[2] LI J, ZHANG J. A study of voice print recognition technology [C]//Proceedings of 2021 International Wireless Communications and Mobile Computing (IWCMC). IEEE, 2021: 1802-1808.

[3] REYNOLDS D A, QUATIERI T F, DUNN R B. Speaker verification using adapted Gaussian mixture models[J]. Digital Signal Processing, 2000, 10(1-3): 19-41.

[4] CAMPBELL W M, STURIM D E, REYNOLDS D A. Support vector machines using GMM supervectors for speaker verification [J]. IEEE Signal Processing Letters, 2006, 13(5): 308-311.

[5] KENNY P, BOULIANNE G, OUELLET P, et al. Joint factor analysis versus eigenchannels in speaker recognition[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(4): 1435-1447.

[6] DEHAK N, KENNY P J, DEHAK R, et al. Front-end factor analysis for speaker verification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 19(4): 788-798.

[7] VERGIN R, O'SHAUGHNESSY D, FARHAT A. Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition[J]. IEEE Transactions on Speech and Audio Processing, 1999, 7(5): 525-532.

[8] YUAN X, LI G, HAN J, et al. Overview of the development of speaker recognition[J]. Journal of Physics: Conference Series. IOP Publishing, 2021, 1827(1): 012125.

[9] SHAN S, LIU J, DUN Y. Prospect of voiceprint recognition based on deep learning[J]. Journal of Physics: Conference Series. IOP Publishing, 2021, 1848(1): 012046.

[10] MOUMIN A A, KUMAR S S. Automatic speaker recognition using deep neural network classifiers[C]//Proceedings of 2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM). IEEE, 2021: 282-286.

[11] BOLES A, RAD P. Voice biometrics: Deep learning-based voiceprint authentication system[C]// Proceedings of 2017 12th System of Systems Engineering Conference (SoSE). IEEE, 2017: 1-6.

[12] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision (ECCV). IEEE, 2018: 3-19.

[13] LANDAU H J. Sampling, data transmission, and the Nyquist rate [J]. Proceedings of the IEEE, 1967, 55(10): 1701-1706.

[14] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[15] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet classification with deep convolutional neural networks [J]. Advances in Neural Information Processing Systems, 2012, 25 (2). DOI:10.1145/3065386

[16] FREE ST Chinese Mandarin Corpus [DB/OL]. 2016. http://www.openslr.org/38/.