生物多样性 BIODIVERSITY SCIENCE

2023, 31 (1): 22308, pages 1–8
doi: 10.17520/biods.2022308

•技术与方法• 中国野生脊椎动物鸣声监测与生物声学研究专题

面向鸟鸣声识别任务的深度学习技术

谢卓钒[1,2,3], 李鼎昭[2,3], 孙海信[2,3*], 张安民[4]

1. 厦门大学电子科学与技术学院(国家示范性微电子学院), 福建厦门 361005; 2. 厦门大学信息学院, 福建厦门 361000; 3. 自然资源部东南沿海海洋信息智能感知与应用重点实验室, 福建厦门 361005; 4. 天津大学海洋科学与技术学院, 天津 300072

摘要: 在生态系统中, 鸟类是重要的组成部分, 对调节生态环境和监测生物多样性至关重要, 甚至可以通过监测鸟群动向与监听鸟群异常鸣声对地震、海啸等自然灾害进行辅助预测和防范, 为此, 鸟鸣声识别和异常鸣声监测成为热门的研究方向。然而, 由于传统鸟鸣声识别方法存在特征提取不充分等问题, 导致识别率不高。本文采用融合特征的方法结合深度学习技术提取鸟鸣声特征, 融合特征选择改良后的对数梅尔谱差分参数同原始信号参数拼接所得的特征; 深度学习方法是基于DenseNet121网络结构, 并融入自注意力模块与中心损失函数进行鸟鸣声识别。自注意力模块部分提高了关键通道的特征表达能力; 中心损失函数可解决类内特征不紧凑问题。我们通过消融实验对比验证, 对在Xeno-Canto世界野生鸟类声音公开数据集上选取的10种鸟类声音进行识别, 准确率达到96.9%。代码已开源至Github: https://github.com/CarrieX6/-Xeno-Canto-.git。

关键词: 鸟鸣声识别; 特征融合; 自注意力模块; 中心损失函数

谢卓钒, 李鼎昭, 孙海信, 张安民 (2023) 面向鸟鸣声识别任务的深度学习技术. 生物多样性, 31, 22308. doi: 10.17520/biods.2022308.
Xie ZF, Li DZ, Sun HX, Zhang AM (2023) Deep learning techniques for bird chirp recognition task. Biodiversity Science, 31, 22308. doi: 10.17520/biods.2022308.

Deep learning techniques for bird chirp recognition task

Zhuofan Xie[1,2,3], Dingzhao Li[2,3], Haixin Sun[2,3*], Anmin Zhang[4]

1 *School of Electronic Science and Engineering (National Model Microelectronics College), Xiamen University, Xiamen, Fujian 361005*
2 *School of Informatics, Xiamen University, Xiamen, Fujian 361000*
3 *Key Laboratory of Southeast Coast Marine Information Intelligent Perception and Application, Ministry of Natural Resources, Xiamen, Fujian 361005*
4 *School of Marine Science and Technology, Tianjin University, Tianjin 300072*

ABSTRACT

Background: In the ecosystem, birds are an important component, which is crucial for regulating the ecological environment and monitoring biodiversity, and can even assist in predicting natural disasters such as earthquakes and tsunamis by monitoring the movement of birds and listening to their abnormal calls, so bird sound recognition and abnormal call detection have become popular research directions. However, low recognition rate is caused to the problems of insufficient feature extraction in traditional bird sound recognition methods.

Method: In this paper, we used a fusion feature method combined with deep learning to extract bird sound features. The fusion features were obtained by splicing the original signal parameters with the modified log-Meier spectral difference parameters; the deep learning method was based on the DenseNet121 network structure and incorporated the self-attention module and the central loss function for bird sound recognition. The self-attentive module partially improved the feature representation of key channels; the central loss function was used to solve the problem of incompact intra-class features. We used the data of 10 bird sounds from the Xeno-Canto World Wild Bird Sounds public dataset to test the accuracy of bird chirp recognition.

Conclusion: In this paper, a neural network structure containing self-attention mechanism and center loss function is proposed for bird song recognition. Its verification accuracy reaches to 96.9%. The code is open source to Github: https://github.com/ CarrieX6/-Xeno-Canto-.git.

收稿日期: 2022-06-08; 接受日期: 2022-07-28
基金项目: 国家自然科学基金(61971362)和福建省自然资源科技创新项目(KY-080000-04-2021-030)
* 通讯作者 Author for correspondence. E-mail: hisensessun@163.com

Key words: bird chirp recognition; feature fusion; self-attentive module; central loss function

目前全球自然界中已知的鸟类超过一万种, 鸟类对生态系统的稳定至关重要(杨俊锋等, 2022)。鸟类位于食物链的上层, 从它们的生存状态可以了解生物圈中的变化, 因此, 鸟类是栖息地质量和环境污染的绝佳指标(Dagan & Izhaki, 2019)。鸟类体型较小, 活动灵敏, 但鸣叫声清晰响亮, 区分度高, 因此选择鸟鸣声进行研究的难度较之图像大大减小。通过适当的声音监测和分类, 可以通过鸟群状态变化感知某一地区生活质量的变化。除了作为环境污染的评价指标外, 鸟声识别在现实中的应用也十分广泛, 因为鸟鸣声包含丰富的生态学信息, 这些信息可以运用到动物行为分析、监护、生态环境的恢复等方面(Incze et al, 2018)。研究鸟声识别技术能够实现长时间无人值守监测, 减轻了生态保护人员的工作, 同时极大节约了成本。收集到的鸟声数据可供我们进行生态监测, 以了解鸟类的分布状态, 分析鸟群迁徙动向。同时, 研究鸟声识别技术对鸟害预防也意义重大。我国对输电线路鸟害防护方面的研究已经持续了多年, 随着人工智能的兴起, 鸟害预防的研究思路从物理被动防护转向了与人工智能技术结合的主动驱赶(Mahendra et al, 2021)。研究人员将输电线路驱鸟器与人工智能技术相结合, 通过鸟鸣声识别技术, 在对输电线路鸟类进行有效监测后启动电子驱鸟器(宋福春等, 2021)。综上, 对鸟鸣声的识别技术进行研究具有重要意义(吕坤朋等, 2021)。

本文基于Xeno-Canto世界野生鸟类声音公开数据集(https://xeno-canto.org), 应用鸟鸣声识别方法对真实生活中采集到的鸟鸣声数据进行分析。因为深度学习任务中的特征提取和分类器设计一样重要, 所以本文研究了鸟鸣声的特征提取以及鸟鸣声识别分类器的选择两方面的内容。在保证特征包含了鸟鸣声数据足够信息的情况下, 对模型进行设计并迭代优化。本文所做主要工作如下: (1)数据分析及预处理。对真实鸟鸣声数据进行分析, 明确了数据集中鸟鸣声的分布情况及不同鸟鸣声时频域的呈现形式。对鸟鸣声数据进行预加重(pre-weighting)、分帧加窗(framing and windowing)等操作。(2)特征融合。对梅尔滤波器进行改进, 剔除原有的离散余弦变换操作, 直接提取对数梅尔谱特征。为避免失去时域上的连续信息, 本文加入一阶差分与二阶差分参数, 并与未经过梅尔提取的原始鸟鸣声信息进行拼接, 得到一个融合特征。(3)设计神经网络架构进行分类。设计一个基于DenseNet121结构的网络架构, 同时引入自注意力模块(self-attentive module)对感兴趣的鸟鸣声进行聚焦, 最后融合softmax损失函数与center loss损失函数对模型进一步优化。

1 方法

本文的识别模型由“特征提取”“特征融合”和“多分类模型”3个关键步骤组成。通过梅尔频率倒谱系数(Mel Frequency Cepstral Coefficents, MFCC)算法进行特征提取, 此过程会损失一部分的信息(Zhang et al, 2021), 所以本文在特征提取的同时也进行了特征融合。多分类模型将特征融合后的特征作为输入, 采用卷积网络模型训练鸟声识别模型。本文的识别模型基本流程图如图1所示。

1.1 数据预处理

若直接对采集到的声信号数据提取特征并分类识别, 信号所带有的环境因素会对特征产生影响, 不能达到期望的效果, 所以需要对数据进行预处理,

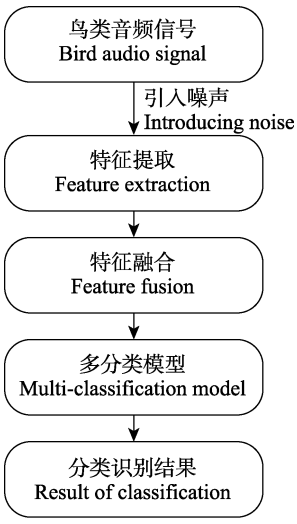

图1 鸟鸣声识别模型流程图
Fig. 1 Flow chart of bird chirp recognition model

使样本数据中的信息更为明显, 减少对特征提取的影响。根据对数据的分析以及对本领域的研究, 本文主要采用预加重、分帧加窗(Petmezas et al, 2022)、幅值规整的方式进行鸟声信号数据预处理。

本文经过Box 1中的步骤对鸟声音频进行预处理, 得到多段时间信号序列(x_1, x_2,…, x_n), 这些序列为特征提取的输入。

Box 1 鸟声数据集处理流程

1. 输入鸟声音频原始的WAV格式文件;

2. 将数据文件切分为5 s每段的音频文件, 采样率为32 kHz;

3. 将音频的信号值以0.97的倍率进行预加重, 计算各采样点的信号值变化;

4. 使用50%重叠的汉明窗对连续采样点分帧加窗, 得到多段近似平稳时间信号序列。

1.1.1 预加重

各类鸟声信号中包含着较为丰富的低频成分, 而高频成分较少。为了得到稳定的信号特征, 需将其高频进行提升。这一方面是由信号本身的特点所决定的, 另一方面主要是由于信号在传播的过程中高频成分衰减率要比低频高。为了加强高频特征在信号特征提取中的作用, 尤其重要的是对远距离传播的信号进行高频提升。设 n 时刻的语音采样信号值为$s(n)$, (n = 1, 2,…, N), 采用(1)式对 $s(n)$ 进行预加重, 预加重后所得的输出信号值为$y(n)$。

$$y(n)=s(n)-us(n-1) \tag{1}$$

其中 u 值为预加重系数, 其大小决定了预加重后音频信号变化的幅度大小。

1.1.2 分帧加窗

数据集中各类别鸟声音频长度长短不等且数量不均, 这使得鸟声识别过程困难重重(Dai et al, 2021)。并且由于音频信号长期呈现出的非平稳性, 使得不能采用一般信号处理的手段来分析。在10–30 ms时间内, 信号可被视为平稳信号。所以可以通过对预加重后的鸟声信号$y(n)$加以分帧来解决这个问题, 分帧可以将一段未知长度的音频数据分割成许多固定长度的片段$y_1(n)$, $y_2(n)$,…, $y_k(n)$, 即取固定值的采样点 n 分割信号。为了保证鸟声音频片段的完整性, 不丧失截断的鸟声信息, 在分帧过程中可使帧与帧之间存在较大的重叠。分帧能够在增加训练样本个数的同时, 避免在分类识别模型训练时出现小样本问题。本文中采用50%重叠的汉明窗(Hamming window)对原始信号进行分帧, 采样率为32 kHz, 每一帧的长度取25 ms, 则每一帧对应的样本数量为800个, 计算公式如下:

$$\text{frame}_{length}=\text{sample}_{rate}\times\text{frame}_{size}=32\times0.025=800 \tag{2}$$

1.2 特征提取

在深度学习领域, 特征的选择和提取发挥着越来越重要的作用, 一个好的特征有时甚至要比一个好的模型更为关键。低层特征包含原始数据中的更多信息, 但是经过的操作少, 包含的噪声信息更多; 高层信息包含噪声少, 但是可能在提取过程中损失掉关键信息。融合不同尺度的特征是一种特征提取的新思路。

1.2.1 梅尔谱系数

特征的具体提取流程和计算过程如下:

(1)预加重、分帧、加窗。确定经过预处理后的鸟声信号每一帧的采样点数n。

(2)时频转换。利用基础的离散傅里叶变换将每帧的时域信号 $s(n)$ 转化为频域以便进行后续的信号分析, 并取转化后信号的模的平方得到离散功率谱。

(3)使用梅尔滤波器组进行过滤。根据信号幅度谱 $S(n)$ 求经过 L 个滤波器的输出, 公式如(1) (3)所示。即通过梅尔滤波器设置方式设置 L 个滤波器, 求 $S(n)$ 和 $W_L(n)$ 的乘积之和, 可以得到 l 个参数 $m(l), l=1,2,...,L-1$。

$$m(l)=\sum_{s=o(l)}^{h(l)}W_L(n)\left|S(n)\right|\,l=1,2,\ldots,L \tag{3}$$

$$W_L(n)=\begin{cases}\dfrac{n-o(l)}{c(l)-o(l)} & o(l)\le s\le c(l)\\ \dfrac{h(l)-n}{h(l)-c(l)} & c(l)\le s\le h(l)\end{cases} \tag{4}$$

(4)对数操作。由于正常人耳对声音的整体感觉并不完全是线性的, 利用对数这种非线性向量关系可以更好地模拟人耳所能得到感知的声音过程。

经过对数操作的梅尔谱系数存在相关性, 为了避免在一些机器学习算法中预先假设数据导致的不相关, 研究者们则提出使用离散余弦变换(discrete cosine transform, DCT)对梅尔谱系数进行压缩得到一组不相关的梅尔频率倒谱系数。但经过复杂的神经网络架构, 在训练的过程中可以将相关

性的问题弱化, 同时DCT作为一种线性变化有可能会损失信号中的一些非线性信息, 所以在本文中舍弃了DCT变换, 从而直接提取梅尔谱系数。梅尔谱系数提取流程如图2所示。

1.2.2 特征融合

由于梅尔谱系数的鲁棒性较差, 为了使特征更能体现时域连续性, 可以在特征维度增加前后帧信息。所以选择加入该特征时间维度上提取的一阶、二阶差分做补充, 与原梅尔谱系数特征一同组成融合特征, 提高对信号噪声的处理效果。差分参数是当前帧的前两帧和后两帧的线性组合。

本文将经过预处理的梅尔谱系数同经过差分的梅尔特征进行拼接(concat), 从而得到一个新的融合特征, 这个特征中既包含了原始信号梅尔谱系数, 使得后续的机器学习模型可以从中直接提取到适合用于分类的特征, 也包含了在语音识别中表现良好的高维对数梅尔谱差分特征。因为该融合特征通过融合方式包含了鸟声数据原始时频信息以及模拟人耳听觉感官得到的信息, 使得它能够从多个角度对鸟鸣声信号数据加以表达。对鸟鸣声信号特征处理提取并融合的全过程如图3所示。假设两个输入特征x和y的维数为p和q, 则输出特征z的维数为$p + q$。

通过预处理和特征提取, 得到了可以直接输入

图2 梅尔谱系数提取流程图

Fig. 2 Mel spectrum coefficient extraction flow chart

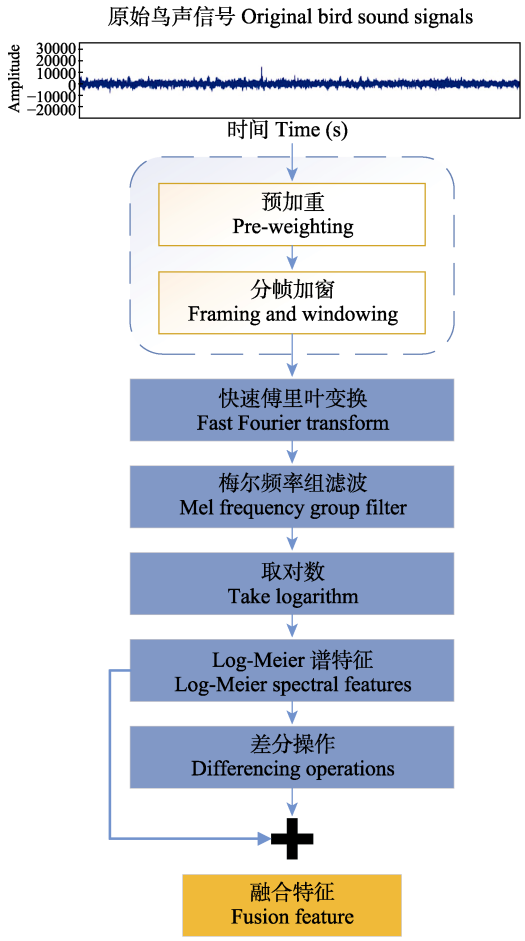

图3 鸟鸣声特征提取与融合流程图。✚为拼接(concat)操作, 即直接将两个特征进行连接。

Fig. 3 Bird chirp feature extraction and fusion flow chart. ✚ means concat operation, which is connecting two features directly.

到分类模型中的三维特征矩阵(Box 2), 矩阵的输入特征为三通道, 使用一阶差分作为第二通道特征, 使用二阶差分作为第三通道特征。

Box 2 融合特征提取流程

输入: 音频时间信号序列组(x_1, x_2,…, x_n)

1. 输入5 s鸟声音频段的时间信号序列帧;

2. 对每帧信号经过短时傅里叶变换得到信号的频谱特征;

3. 计算经过60阶的梅尔频率组的输出的对数能量和;

4. 计算一阶差分、二阶差分;

5. 将梅尔谱特征、梅尔一阶差分值与梅尔二阶差分值拼接组成融合特征。

输出: 三维特征矩阵[$\boldsymbol{X}_i$, $\boldsymbol{Y}_i$, $\boldsymbol{Z}_i$]

1.3 神经网络架构设计

1.3.1 网络结构

随着深度学习的发展, 不同的卷积神经网络(convolutional neural networks, CNN)被提出, 如AlexNet (Krizhevsky et al, 2017)、VGG (Simonyan & Zisserman, 2014)、ResNet (He et al, 2016)等。随着深度学习技术的发展, 增加网络层数成为了优化网络模型最直接的手段, 但对于非海量的数据样本, 过深的网络模型反而会造成梯度消失, 从而导致更低的识别精度。因此本文提出了一种结合自注意力机制(self-attentive)和DenseNet网络模型的鸟声分类方法。

DenseNet是一种卷积神经网络结构(Huang et al, 2017), 主要包含卷积层(convolutional layer)、密集块(dense block)、过渡层(transition layer)和分类器(classifier)。它采用的是密集连接机制, 以前馈(feedforward)方式与各个密集块中的层直接连接。它把前面所有层特征图的输出均作为当前层的输入。密集连接的方式使得网络能够同时挖掘空间域与时间域的特征, 并且所使用的注意力机制能够帮助网络聚焦重要特征, 减少无关信息的干扰。由此解决梯度消失的问题, 提升所提框架性能。本文采用DenseNet121模型, 该模型总共包括4个密集块和121个层。网络中的每一层均以紧密连接的方式与随后的所有层相连, 最后是将softmax loss和center loss结合对模型进一步优化。如图4为Attention-DenseNet的网络结构。将数据X_0输入卷积网络, 网络由l层组合而成, H_l对上一层传入的数据进行线性变换, 包括批量归一化(batch normalization, BN), ReLu激活函数, 池化或卷积操作, 令第l层的输出为X_l。

1.3.2 自注意力模块

在一段音频中所包含的声音可能包含背景噪声、风声等非兴趣噪声, 为了让网络关注所感兴趣的鸟声部分, 此处引入注意力模块进行兴趣区域聚焦。

本文使用自注意机制筛选前两层输出的特征图的通道特征, 用来提升特定通道的重要特征表达能力, 聚焦到网络的兴趣区域。该设计思路是基于Buades等人提出的非局部均值(non-local means, NLM)降噪算法(Buades et al, 2011)。自注意力模块结构图如图5所示。

首先, 通过3条分支分别使用1 × 1卷积对输入的特征图进行通道压缩, 使其通道维度压缩为一维, 由此减少无关信息冗余, 提高后续计算速度。然后, 在F(X)分支输出的特征图上进行转置操作, 并将结果转置后的结果与g(X)分支的输出特征进行矩阵相乘, 接着使用softmax对其进行归一化。在这一过程中, 向量间的余弦相似度通过矩阵乘积进行表征, 乘积的结果即表示为特征图间的相似度。最后, h(X)分支的特征图与归一化后的注意力矩阵进行相乘, 得到最后的特征图。根据类内相似度在通道维度上进行权重的加权重分配, 接着将结果输入softmax函数中, 并且使用1 × 1卷积核对输出结果进行处理, 使其通道数与输入特征图的通道保持一致。通过这种方式, 输出的特征经过注意力机制进行了权重重分配, 使其能够充分表达。

最后, 为了优化系统在高维特征检索时的时间开销(time consuming)及存储上的空间开销(space consuming), 本文采用主成分分析法(PCA)对高维特

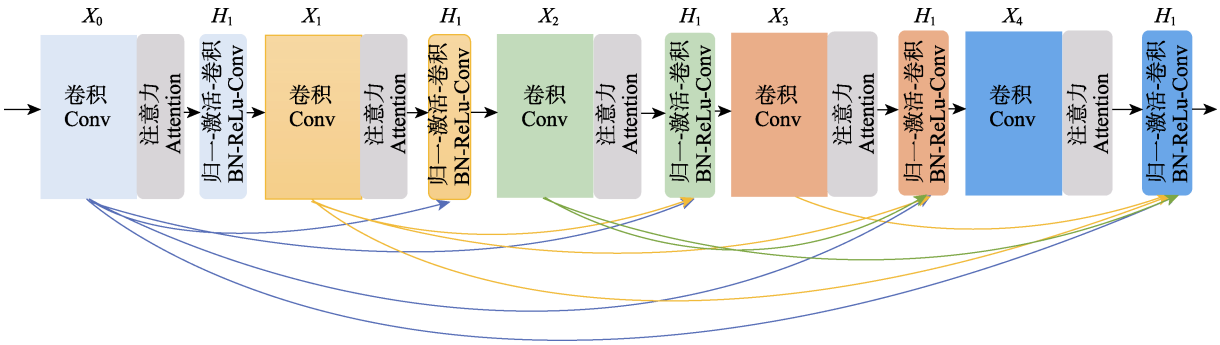

图4 Attention-DenseNet网络结构图。Conv: 卷积层; BN: 批量归一化; ReLu: ReLu激活函数。
Fig. 4 Attention-DenseNet structure diagram. Conv, Convolutional layer; BN, Batch normalization; ReLu, ReLu activation function.

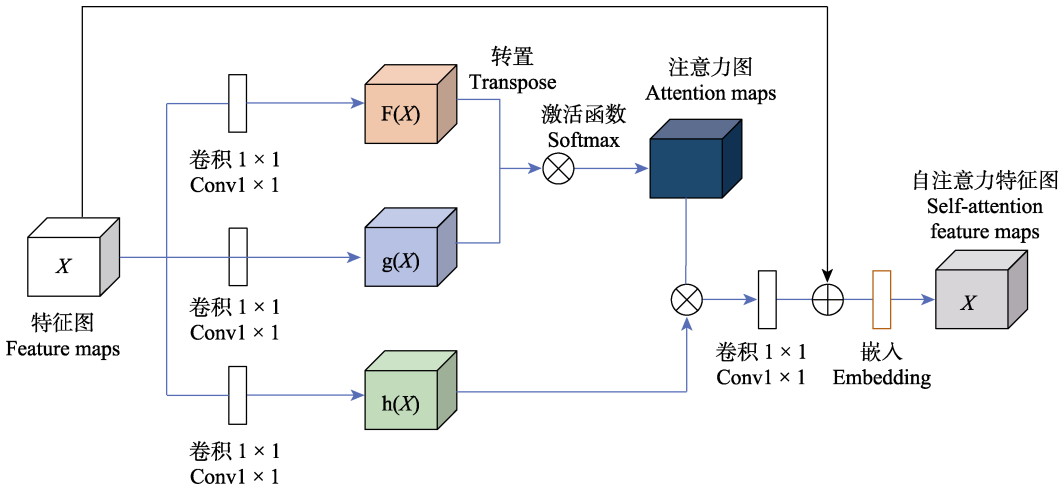

图5　自注意力模块结构图
Fig. 5　Self-attention module structure diagram

征进行降维，最终输出能够高效表征图像特征的128维特征向量。

1.3.3　中心损失函数

在进行音频采集时，同类别的鸟鸣声数据会因其采集方式、环境因素、个体特征等易出现较大的差距，从而导致在类间识别时由于类内特征过于分散而相互混杂。故而本文使用了两种损失函数相结合的方法来改善上述问题。损失函数(L)计算公式如下:

$$L=\mathcal{L}_S+v\mathcal{L}_C \tag{5}$$

$$\mathcal{L}_S=-\sum_{i}^{m}\log\frac{e^{W_{y_i}^{T}x_i+b_{y_i}}}{\sum_{j=1}^{n}e^{W_{y_j}^{T}x_j+b_{y_j}}} \tag{6}$$

$$\mathcal{L}_C=\frac{v}{2}\sum_{i=1}^{m}\|x_i-C_{y_i}\|_2^2 \tag{7}$$

为了解决类内特征不紧凑的问题，本文引入了中心损失，使用超参数公式v控制 softmax loss 和 center loss 的比重，本实验中v初始化为1。公式(6)中 $\mathcal{L}_S$ 为softmax loss 的结果，x_i 表示网络提取到的样本特征，$W_j\in\mathbb{R}^d$ 表示权重矩阵$W\in\mathbb{R}^{d\times n}$ 的第j列，m表示mini-batch包含的样本数量，n表示类别数。中心损失 $\mathcal{L}_C$ 如公式(7)所示，C_{yi}表示第yi类别的特征中心，使用该损失函数是为了减小同一类别yi中样本的差异性，拉近第yi类别中样本与特征中心的距离，并通过反向传播传递训练后的损失值，以达到优化网络的目的。

2　案例分析

2.1　数据集获取

本文实验所采用的模型训练鸟类音频文件数据集全部来自Xeno-Canto世界野生鸟类声音公开数据集(https://xeno-canto.org/)。该网站共收录了世界各地10,000多种鸟类的声音，其中亚种的包含5,000多种，采集地点包含森林、草地、湿地等。本文随机选取了10种鸣叫声单调的鸟种音频，避免了鸣唱复杂的雀形目鸟类，表1中列举了所使用的鸟类学名、时长等信息，由于每一种鸟类的音频采集地点和设备都有差异，此处标注了每类鸟种的数据集网址，本实验中使用的所有数据皆下载自对应的网址。为避免数据不平衡问题，在下载数据时进行了样本量统一。通过对每个鸟类音频进行切分，统一长度为5 s一个的WAV文件，然后使用(Box 2)所述算法，最终生成45,760个样本。本文划分的训练集和测试集比例为7∶3。

3.2　训练结果分析

本实验在ubuntu操作系统中运行，使用显卡为NVIDIA-3080、并配备32G内存。编译器为PyCharm，并在其中使用Pytorch框架完成神经网络架构设计。本实验以多种鸟类声音分类为研究背景，对测试集的13,728个样本、10个类别进行分类识别。使用基于注意力机制的DenseNet模型进行实验，模型使用的参数见表2。

(1)特征效果对比分析。融合特征是将改良后的

表1　本研究所采用的模型训数据集说明

Table 1　Dataset description of model training dataset used in this study

鸟类中文名称 Chinese name	鸟类英文学名 Latin name	库内标签 Registered label	样本时长 Sample length (s)	数据来源 Data source
美洲麻鸦	*Botaurus lentiginosus*	amebit	23,960.7	https://xeno-canto.org/species/Botaurus-lentiginosus
白头海雕	*Haliaeetus leucocephalus*	baleag	22,744.8	https://xeno-canto.org/species/Haliaeetus-leucocephalus
布氏雀鹀	*Spizella breweri*	brespa	23,880.6	https://xeno-canto.org/species/Spizella-breweri
普通拟八哥	*Quiscalus quiscula*	comgra	23,517.5	https://xeno-canto.org/species/Quiscalus-quiscula
角鸬鹚	*Phalacrocorax auritus*	doccor	22,183.7	https://xeno-canto.org/species/Phalacrocorax-auritus
灰斑鸠	*Streptopelia decaocto*	eucdov	22,837.2	https://xeno-canto.org/species/Streptopelia-decaocto
长嘴啄木鸟	*Leuconotopicus villosus*	haiwoo	23,009.2	https://xeno-canto.org/species/Leuconotopicus-villosus
暗背金翅雀	*Spinus psaltria*	lesgol	22,521.4	https://xeno-canto.org/species/Spinus-psaltria
环颈潜鸭	*Aythya collaris*	rinduc	21,653.2	https://xeno-canto.org/species/Aythya-collaris
白喉雨燕	*Aeronautes saxatalis*	whtswi	22,491.7	https://xeno-canto.org/species/Aeronautes-saxatalis

对数梅尔谱差分特征同原始信号拼接所得的特征。分别利用原始特征、对数梅尔谱差分特征和融合特征在VGG11、ResNet18和DensNet121模型中进行对比实验，测试特征效果。测试结果见表3，表中显示了融合特征在Densnet121和VGG11模型下准确率最高。

表3中可以看出融合特征在VGG11、ResNet18和Densnet121模型的准确率都在93%以上，虽然准确率相差不大，但是在参数方面DensNet121的总参数远远小于其他两个模型，从综合考虑DensNet121具有空间复杂度低，为此本文选择DensNet121作为基础模型。

(2)消融实验分析。为了体现本文所提到改进点的有效性，本文通过消融实验对比，并对结果进行分析。通过对比改进点在3种经典网络框架上的应用结果，选取最适用于针对本文数据的框架结构。具体为以下三点：一是原始特征和对数梅尔谱差分特征进行融合特征；二是加入自注意力模块，强化音频信号中关键细节特征的表达能力；三是使用softmax loss和center loss结合优化，在加大类间距离的同时，缩短类内距离，使样本特征在特征空间的分布更加合理。具体实验结果如表4，本文使用的方法在测试集数下准确率达到了96.9%。

本文提出的模型对鸟声数据集提取得到的DensNet121 + 融合特征 + 注意力机制 + 中心损失函数测试过程的性能变化曲线，如图6所示。

从图6可以看出该模型在前10轮次(Epoch)快速达到训练最优解，收敛速度变慢趋于平稳，最终在训练集上的准确率达到96.9%，损失接近于1%。

表2　DenseNet模型参数列表

Table 2　DenseNet model parameter list

参数名称　Parameter name	参数值　Parameter value
批大小　Batch_size	256
时期数　Epochs	50
学习率　Learning rate	0.001
优化器　Optimizer	自适应矩估计优化器　Adam optimizer
损失函数　Loss function	分类交叉熵　Categorical_cross-entropy

表3　VGG11、ResNet18与DensNet121采用不同特征准确率对比。加粗数值为使用融合特征所得的准确率。

Table 3　Comparison among different feature accuracies of VGG11, ResNet18 and DensNet121. Bold value is the accuracy calculated by the fusion feature.

特征提取方法 Feature extraction method	准确率 Accuracy	总参数量 No. of parameters
VGG11 + 原始特征 VGG11 + Original feature	0.906	1.38e8
VGG11 + 对数梅尔谱差分特征 VGG11 + Log-Meier spectral differential characteristics	0.926	1.38e8
VGG11 + 融合特征 VGG11 + Fusion feature	**0.935**	1.38e8
ResNet18 + 原始特征 ResNet18 + Original feature	0.896	1.11e7
ResNet18 + 对数梅尔谱差分特征 ResNet18 + Log-Meier spectral differential characteristics	0.912	1.11e7
ResNet18 + 融合特征 ResNet18 + Fusion feature	**0.933**	1.11e7
DensNet121 + 原始特征 DensNet121 + Original feature	0.901	6.94e6
DensNet121 + 对数梅尔谱差分特征 DensNet121 + Log-Meier spectral differential characteristics	0.932	6.96e6
DensNet121 + 融合特征 DensNet121 + Fusion feature	**0.939**	6.96e6

表4　基于DenseNet121的对比实验

Table 4　Comparison experiment based on DenseNet121

模型 Model	准确率 Accuracy
DensNet121 + 融合特征 DensNet121 + Fusion feature	0.939
DensNet121 + 融合特征 + 注意力机制 DenseNet121 + Fusion feature + Attention	0.953
DensNet121 + 融合特征 + 注意力机制 + 中心损失函数 DenseNet121 + Fusion feature + Attention + Center loss function	0.969

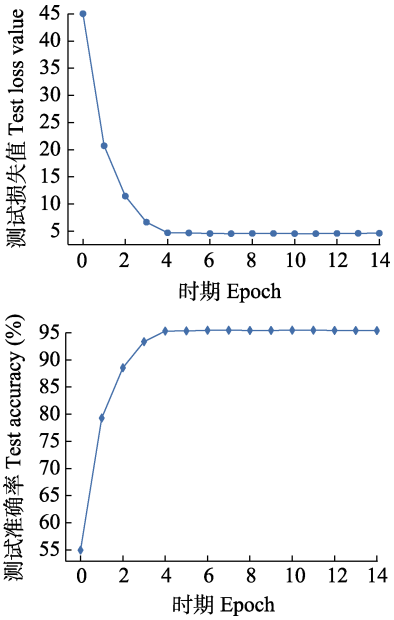

图6　实验测试损失值与识别准确率。上图为0–14时期损失值点图，下图为0–14时期准确率点图。

Fig. 6　Experimental test loss value and recognition accuracy. The above figure is the dot plot for the loss values of 0–14 epochs, and the figure below is the dot plot for the accuracies of 0–14 epochs.

3　结论

本文从特征提取技术和设计模型对所做主要工作加以验证和评估，并通过对比实验结果进行评估分析，实现鸟声信号的目标识别。同时验证了改良后的对数梅尔谱差分参数同原始信号参数拼接所得的特征方法，该方法有效地提升了鸟声目标分类的准确率。最后，评估结果在构造的数据集上进行测试，在测试集上准确率达到96.9%，实验验证了所提出模型的有效性与可行性。所有代码已开源至Github: https://github.com/CarrieX6/-Xeno-Canto-.git。

ORCID

谢卓钒 https://orcid.org/0000-0001-5297-0701

参考文献

Buades A, Coll B, Morel JM (2011) Non-local means denoising. Image Processing on Line, 1, 208–212.

Dagan U, Izhaki I (2019) Understory vegetation in planted pine forests governs bird community composition and diversity in the eastern Mediterranean region. Forest Ecosystems, 6, 29.

Dai YS, Yang J, Dong YW, Zou HP, Hu MZ, Wang B (2021) Blind source separation-based IVA-Xception model for bird sound recognition in complex acoustic environments. Electronics Letters, 57, 454–456.

He KM, Zhang XY, Ren SQ, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. Las Vegas, NV, USA.

Huang G, Liu Z, van der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269. Honolulu, HI, USA.

Incze Á, Jancsó HB, Szilágyi Z, Farkas A, Sulyok C (2018) Bird sound recognition using a convolutional neural network. In: 2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY), pp. 295–300. Subotica, Serbia.

Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60, 84–90.

Lü KP, Sun B, Zhao YX (2021) Research on bird recognition method based on bird singing and deep learning. Bulletin of Science and Technology, 37(10), 24–30, 37. (in Chinese) [吕坤朋, 孙斌, 赵玉晓 (2021) 基于鸟鸣声及深度学习的鸟类识别方法研究. 科技通报, 37(10), 24–30, 37.]

Mahendra M, Nasution MA, Rahmayanti F, Islama D (2021) Application of appropriate technology for automatic bird pest removal and automatic fish feed in the Minapadi system in Beutong Nagan Raya District. International Journal of Community Service, 1(3), 231–237.

Petmezas G, Cheimariotis GA, Stefanopoulos L, Rocha B, Paiva RP, Katsaggelos AK, Maglaveras N (2022) Automated lung sound classification using a hybrid CNN-LSTM network and focal loss function. Sensors, 22(3), 1232.

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition, doi: arXiv: 1409.1556.

Song FC, Ding XM, Yao F, Rui SJ, Chen R (2021) Research on railway intelligent bird repellent based on sensor technology and Internet of Things technology. Railway Engineering Technology and Economy, 36(1), 33–37. (in Chinese) [宋福春, 丁小明, 姚发, 芮胜骏, 陈容 (2021) 基于传感器技术和物联网技术的铁路智能驱鸟器的研究. 铁路工程技术与经济, 36(1), 33–37.]

Yang JF, Liu QQ, Zhang K, Lin QQ, Hou JH (2022) Diversity of bird community in spring in Bodhi Islands, Hebei Province. Journal of Hebei University (Natural Science Edition), 42, 182–189. (in Chinese with English abstract) [杨俊锋, 刘琪琪, 张侃, 林庆乾, 侯建华 (2022) 河北菩提岛诸岛春季鸟类群落多样性. 河北大学学报(自然科学版), 42, 182–189.]

Zhang Y, Zeng JF, Li YM, Chen D (2021) Convolutional neural network-gated recurrent unit neural network with feature fusion for environmental sound classification. Automatic Control and Computer Sciences, 55, 311–318.

(责任编委: 肖治术　责任编辑: 李会丽)