李怀城,杨道武,温治芳,等.基于Inception-CSA深度学习模型的鸟鸣分类[J].华中农业大学学报,2023,42(3):97-104.
DOI:10.13300/j.cnki.hnlkxb.2023.03.012

基于Inception-CSA深度学习模型的鸟鸣分类

李怀城,杨道武,温治芳,王亚楠,陈爱斌

中南林业科技大学计算机与信息工程学院/人工智能应用研究所,长沙 410004

摘要 为进一步提高通过声音识别鸟类的精确度,本研究提出基于Inception-CSA深度学习模型的鸟鸣声分类方法,包含鸟鸣声音频样本预处理、特征提取、分类器分类等步骤。首先将鸟鸣声样本预处理成尺寸相同的梅尔频谱图,作为鸟鸣声特征图;其次利用Inception-CSA模型对鸟鸣声特征图进行特征提取,其中Inception模块提取鸟鸣声特征图中的多尺度局部时频域特征,CSA模块获取鸟鸣声特征图的全局注意力权重,将二者的输出结合得到更强的特征图,再次利用最大池化层对特征图进行下采样;最后利用全连接层进行分类,得到最终的分类结果。以采集的华南地区自然环境中的10种野生鸟类的鸣叫声构建数据集,用于实验部分以验证方法的有效性。结果表明,本研究提出的方法在自建数据集上准确率达到了93.11%,相比于基于其他经典模型的分类方法,基于Inception-CSA模型的分类方法在拥有较少模型参数量的同时达到了更高的准确率。

关键词 卷积神经网络;鸟鸣声分类;深度学习;梅尔频谱图;Inception

中图分类号 TP183 **文献标识码** A **文章编号** 1000-2421(2023)03-0097-08

随着工业社会的发展,生态环境的保护与修复逐渐成为研究重点[1]。在生态环境中,鸟类是野生动物中最具代表性的类群之一[2],其对栖息地环境变化的反应极为敏感。鸟类物种多样性的组成、数量、生活习性等特征反映了其栖息地生态环境的适宜性、人类社会的发展对该生态系统的影响程度等[3-4]。识别鸟类所属种群对珍稀鸟类的保护以及生态环境检测都有着重要意义,因此,研究一种高精度和强泛化性的鸟鸣声分类方法是很有必要的。目前,鸟类识别任务主要基于图像信号和音频信号。利用图像信号对鸟类进行分类研究存在一定的局限性[5]。在自然界中,鸟类的活动区间很大,可能位于树梢、灌木丛等地方,增加了图像采集的难度,且图像的光照强弱、清晰度、复杂的背景信息都会影响分类结果。基于鸟鸣声音频信号进行的分类研究,其音频样本的采集具有识别范围广、无遮拦、成本低等优点[6]。但自然环境下的鸟鸣声往往伴随着动物叫声、水流声和风声等环境噪声,给基于鸟鸣声识别的研究带来巨大的挑战[7]。

早期鸟鸣声分类研究主要以传统机器学习的方法为主,包括支持向量机[8]、决策树[8]、随机森林[9]、高斯混合模型[10-11]、隐马尔科夫模型[12]等。该类方法提取特征的能力有限,很难提取到声音特征的时间频率变化,难以应对真实环境下有背景噪声影响的鸟鸣声分类研究。近年来,深度学习模型被证明比传统的机器学习方法更适用于复杂的分类问题[13]。音频样本中的时域信号经过短时傅里叶变换(short time Fourier transform,STFT)等操作可以转换成同时具备时域特征和频域特征的语谱图(spectrogram),利用语谱图能进一步提取音频特征用于音频分类任务。随着深度学习的发展,卷积神经网络(convolutional neural networks, CNN)被证实有强大的特征提取能力,可以用于提取时频谱图的时频域特征,并用于鸟鸣声分类任务[14]。Sprengel等[15]借助短时距傅里叶变换将鸟鸣声音频样本转换成语谱图,并进行归一化处理,然后利用5层单种大小卷积核的卷积神经网络对语谱图进行分类任务,在复杂音频背景下分类效果的平均精度均值为0.69。Joly等[16]将鸟鸣声预处理为时频谱图,并利用ResNet50神经网络模型对时频谱图进行特征提取以及鸟鸣声

收稿日期:2022-09-19
基金项目:国家自然科学基金项目(62276276);智慧物流技术湖南省重点实验室项目(2019TP1015);湖南省研究生科研创新项目(CX20210879)
李怀城,E-mail:Refrain_lhc@163.com
通信作者:陈爱斌,E-mail:hotaibin@163.com

分类任务，最终获得比较好的效果。Anand等[2]将鸟鸣声样本转换成语谱图，再利用卷积层与池化层相结合的神经网络模型对其进行分类，最终在1个包含5种鸟类的鸟鸣声数据集上得到了90%的分类精度。以上的研究都是将鸟鸣声样本转换变成同时包含时域和频域特征的时频谱图，再利用卷积神经网络模型对时频谱图进行分类。这样处理虽然能同时利用到时频域特征，但对于鸟鸣声样本转换得到的特征图，不同鸟类的鸣叫声在时频谱图上的表现也各不相同，同时单种大小卷积核的卷积层能提取到的特征信息相对有限。Szegedy等[17]提出了1种Inception神经网络结构，每次卷积操作包括多种大小卷积核的卷积层，能从特征图中提取到不同感受野下的特征，进而获得更多的特征用于后续的任务。但卷积神经网络的感受野受到卷积核大小和网络深度的影响，导致其难以关注到特征图上远距离特征，因此，单纯的卷积神经网络在特征提取上还存在一定的不足。Hou等[18]提出协调注意力(coordinate attention，CA)，先将二维的特征图利用平均池化压缩成2个维度上的特征向量，然后利用归一化、激活等方式分别得到2个维度上的注意权重向量，再经过矩阵乘法得到特征图全局上的注意权重。其中，注意权重的作用是给予特征图中重要特征部位更高的权重，以此让神经网络模型加强对该部位的关注。音频特征图同时具备时域和频域特征，协调注意力正适用于此，能弥补卷积神经网络不能捕获远距离特征的缺陷。

因此，本研究基于Inception结构和协调注意力，提出Inception-CSA神经网络模型用于鸟鸣声分类任务，旨在为识别鸟类所属种群、保护珍稀鸟类以及检测生态环境质量提供核心算法理论参考。

1 材料与方法

基于Inception-CSA神经网络模型的鸟鸣声分类方法总体结构图如图1所示。首先，对于鸟鸣声样本，经过预加重、分帧、加窗、短时傅里叶变换等操作预处理为基于人耳听觉系统的梅尔频谱图。再利用Inception-CSA模型从梅尔频谱图中提取融合多尺度局部特征和全局注意力权重的鸟鸣声特征图。其中，Inception-CSA模型由原生Inception模型利用sin函数和协调注意力模块改进得到。接着利用全连接层对特征图进行分类，最终得到每个类别的概率，取最大概率的类别作为最终的分类结果。

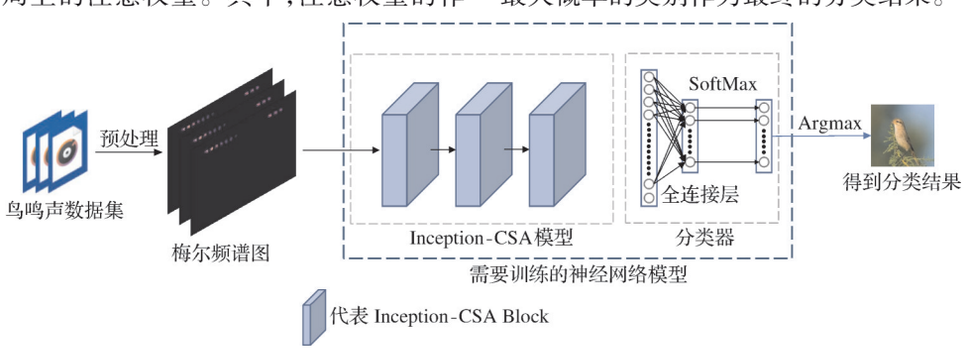


图1 基于Inception-CSA模型的鸟鸣声分类方法总体结构图

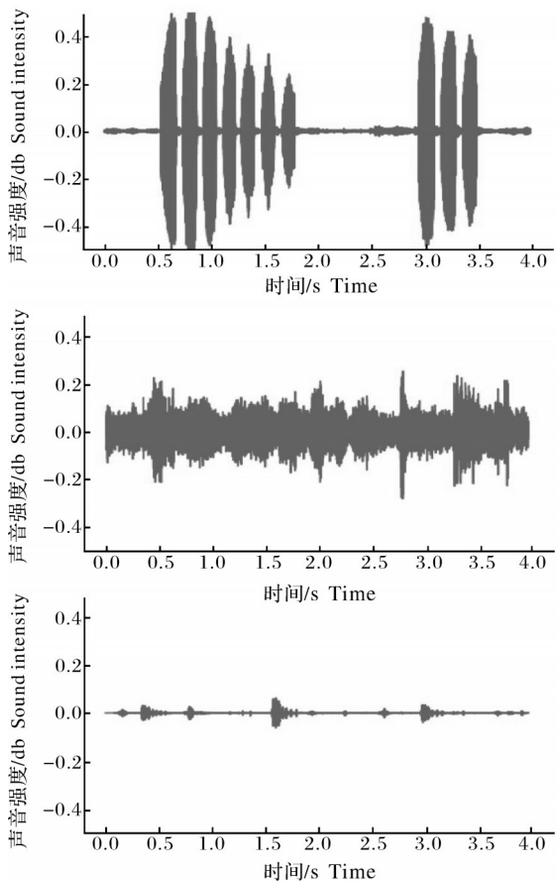
Fig.1 Overall structure of bird song classification method based on Inception-CSA model

1.1 梅尔频谱图特征提取

自然环境中的鸟鸣声在环境噪声、鸣叫声响度等因素的影响下，表现出来的波形图差别较大，甚至同种鸟类的雏鸟和成年鸟的鸣叫声也存在明显的差异。图2为银喉长尾山雀(*Aegithalos glaucogularis*)在不同情况下的鸣叫声波形图。由图2可知，噪声的存在会遮挡鸟鸣声的波形，鸟鸣声响度较小时波形不明显，增加了基于波形图进行分类任务的难度。

对于1个时长为4 s的音频，以22 050 Hz采样频率进行采样处理，会得到1个包含大约80 000个采样点的信号序列，这些采样点仅包含时域特征，且数量极多杂乱无章，不易从中提取特征进而进行分类任务。基于上述限制，本研究将音频样本可视化，即将鸟鸣声样本转换成同时包含时域特征和频域特征的基于人耳听觉的梅尔频谱图，将梅尔频谱图作为Inception-CSA模型的输入特征图进行分类任务。

梅尔频谱图的提取过程如下：首先将原始音频进行预加重、分帧、加窗等操作，再进行傅里叶变换得到频谱图。傅里叶变换主要从频域突出音频特征，公式如式(1)所示，图2中颜色越深的地方，则代表该帧级区域的频率越大。



A：无噪声时的正常鸟鸣声 Normal birdsong when there is no noise；
B：较大噪声时的正常鸟鸣声 Normal birdsong in case of loud noise；
C：无噪声且较小鸟鸣声 No noise and birdsong.

图 2　银喉长尾山雀在不同情况下的鸣叫声波形图

Fig.2　Waveforms of calls of the silver-throated long-tailed tit in different situations

$$x(f,\tau)=\int_{-\infty}^{+\infty} w(t-\tau)y(t)e^{-j2\pi ft}\mathrm{d}t \tag{1}$$

其中，f代表频率，τ代表帧长，对于1个4 s的音频，可以获得174帧，则$\tau\in[0,\ 174)$。$y(t)$为时域信号，x代表频域信号，$w(t-\tau)$是中心位置位于τ的汉明窗(Hamming window)，窗口长度设为2 048，步长为512，采样率为22 050 Hz。对于频谱图利用梅尔滤波器组过滤得到梅尔频谱图为：

$$f_{\mathrm{mel}}=2595\lg\left(1+\frac{f}{700}\right) \tag{2}$$

其中，f代表正常频率，f_{mel}是经过滤波后的梅尔标度频率。因为频域信号有很多冗余，滤波器组可以对频域的幅值进行精简，每个频段用1个值来表示。本研究设置滤波器组个数为64，即$f_{\mathrm{mel}}\in[0,64)$。

1.2　Inception-CSA模型

本研究在Inception模型和协调注意力的基础上提出Inception-CSA模型，其由3个相同的Inception-CSA Block构成，Inception-CSA Block的结构见图3。Inception-CSA Block由2个部分组成，第一部分是多尺度卷积模块，第二部分是改进后的协调注意力模块(coordinate-sin attention，CSA)。

对于1个尺寸为$C\times H\times W$的音频特征图$f\in F^{C\times H\times W}$，$C$代表通道数，$H$代表特征图的频率维度，$W$代表特征图的时间维度。在多尺度卷积模块部分中，为了扩充特征的多样性，特征图分别经过1×1、3×3、5×5共3种大小卷积核的卷积层，提取其不同感受野下的局部时频域特征，得到3种大小均为$\frac{C}{3}\times$

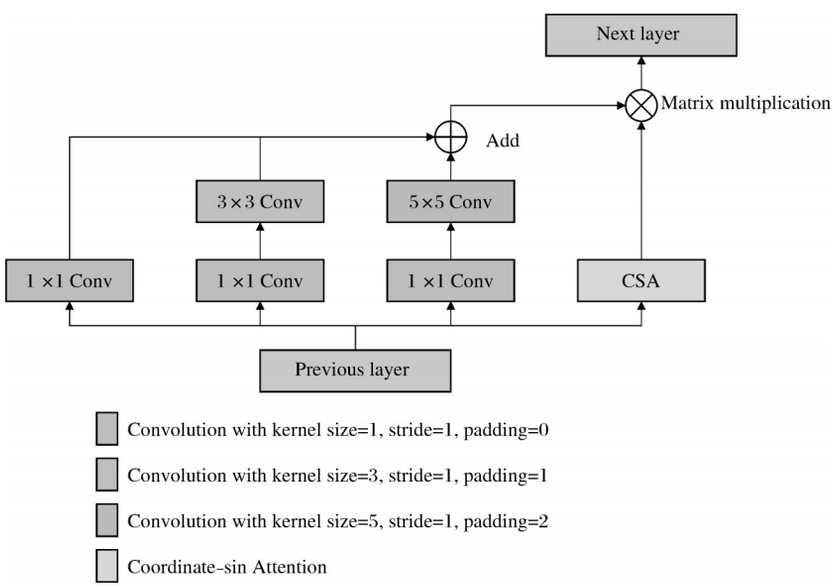


图 3　Inception-CSA Block的结构图

Fig.3　Structure diagram of Inception-CSA Block

$H\times W$ 特征图 $f_1\in F_1^{\frac{C}{3}\times H\times W}$、$f_3\in F_3^{\frac{C}{3}\times H\times W}$、$f_5\in F_5^{\frac{C}{3}\times H\times W}$。其中，$F_1$、$F_3$、$F_5$每个通道上的特征图都能单独代表原始特征图的特征，将3个特征图在通道维度上进行拼接，得到大小为$C\times H\times W$的特征图$f'\in F_{\{1,3,5\}}^{C\times H\times W}$。

多尺度卷积模块虽然能提取不同范围内的局部高维特征，但受卷积层感受野大小的限制，在提取局部特征的同时不能关注到特征图中距离较远的节点的特征，这就导致全局特征的丢失。而且音频包含许多冗杂的信息，利用CSA模块旨在得到特征图中鸟鸣声部分的全局注意权重，让整个网络在特征提取的过程中在全局上更加关注特征图中鸟鸣声部分，CSA模块的结构图见图4。

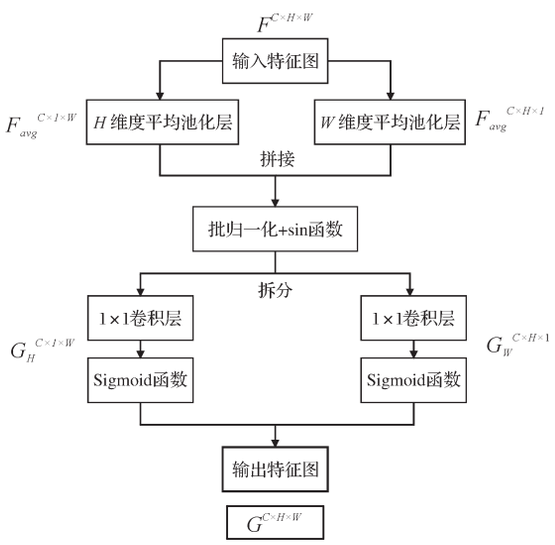


图4 CSA模块的结构图

Fig.4 Structure diagram of CSA module

CSA模块将特征图$f\in F^{C\times H\times W}$分别进行时域和频域2个维度上的平均池化，得到2个维度上的特征向量$f^H\in F_{\text{avg}}^{C\times 1\times W}$、$f^W\in F_{\text{avg}}^{C\times H\times 1}$。接着对特征向量进行归一化，利用sin函数对特征向量进行激活，再经过1×1大小卷积核的卷积层与Sigmoid层，增加注意力模块的非线性映射，得到2个维度上的特征注意权重$g^H\in {G_h}^{C\times 1\times W}$和$g^W\in {G_w}^{C\times H\times 1}$。最后将2个维度上的特征注意权重进行矩阵乘法，得到全局特征注意权重$g\in G^{C\times H\times W}$，见式(3)～(5)。

$$G_W{}^{C\times H\times 1}=\sigma(\text{Conv}(\sin(\text{BN}(\text{AvgPool\_}H(F^{C\times H\times W})))))\tag{3}$$

$$G_W{}^{C\times H\times 1}=\sigma(\text{Conv}(\sin(\text{BN}(\text{AvgPool\_}W(F^{C\times H\times W})))))\tag{4}$$

$$G^{C\times H\times W}=G_H{}^{C\times 1\times W}\times G_W{}^{C\times H\times 1}\tag{5}$$

其中，σ表示Sigmoid函数，Conv表示卷积操作，BN表示归一化操作，sin为周期正弦函数，其可以解决特征值差异悬殊引起的特征注意权重差异过大问题，以便关注权重能更容易集中到特征图中鸟鸣声部分。声音信号经过傅里叶变换，被分解成若干不同频率不同强度正弦波的叠加，以此得到的梅尔频谱图中还留存着正弦波的特征。sin函数具有周期性，其值域为[－1，1]，对特征向量具有一定的约束性，一定程度上也能避免特征之间差异过大引起的过拟合现象。为避免在利用Inception-CSA Block模块对特征图进行下采样的过程中特征图的这些特征丢失，在coordinate-sin attention模块中加入sin函数对特征向量进行激活。sin函数的任意阶导数均为三角函数，不会增加神经网络训练中反向传播的计算复杂度。虽然傅里叶变化的结果能用cos函数表示，其值域与sin函数相同且具有周期性，但cos函数并不适用于此，原因是在数据预处理阶段对梅尔频谱图进行了0填充，该0填充的部分为无声部分，其在下采样中不应该改变，而使用cos函数进行激活，会改变该部分的特征值，从而导致声音特征中引入噪声。

将多尺度卷积模块的输出特征图$f'\in F_{\{1,3,5\}}^{C\times H\times W}$与CSA模块的输出全局特征注意权重$g\in G^{C\times H\times W}$进行矩阵点乘，得到同时具备多尺度局部特征和全局关注的特征图$\bar{f}\in \bar{F}^{C\times H\times W}$，以利用全局特征关注权重让特征图中鸟鸣声部分的特征更加突出。最后再利用3×3池化层对特征图$\bar{f}\in \bar{F}^{C\times H\times W}$进行下采样，得到Inception-CSA Block模块特征最终的提取结果$f\in F^{C\times\frac{H}{2}\times\frac{W}{2}}$。使用池化层能提取鸟鸣声特征图中的纹理特征，以此可以减少噪声的干扰，同时还能将特征图缩小，减少后续网络结构中的计算量。

1.3 分类器模型

梅尔频谱图经过Inception-CSA模型的特征提取得到鸟鸣声特征图，利用全连接层对鸟鸣声特征图进行分类感知，得到10个类别的概率，其中最高概率所对应的类别即为分类的结果。

1.4 试验数据处理

本研究采集了华南区域自然环境中常见的10种野生鸟类的鸣叫声，包括银喉长尾山雀(*Aegithalos glaucogularis*)、黑翅雀鹎(*Aegithina tiphia*)、绿头鸭(*Anas platyrhynchos*)、小白鹭(*Egretta garzetta*)、噪鹃(*Eudynamys scolopaceus*)、家燕(*Hirundo rustica*)、

红耳鸭（*Malacorhynchus membranaceus*）、白鹡鸰（*Motacilla alba*）、珠颈斑鸠（*Streptopelia chinensis*）、暗绿绣眼鸟（*Zosterops japonicus*）。由于这些音频样本的时间长度相差较大，为了使输送进网络的数据标准化，本研究将这些鸟鸣声的音频样本通过音频剪辑统一处理成时长 4 s 左右、采样率为 41 000 Hz 的音频样本，构建了 1 个用于鸟鸣声分类任务的数据集。各个类别的音频样本数量分别为银喉长尾山雀 847、黑翅雀鹎 844、绿头鸭 850、小白鹭 846、噪鹃 852、家燕 930、红耳鸭 919、白鹡鸰 904、珠颈斑鸠 847、暗绿绣眼鸟 926。

试验中将数据集的每个类别随机选取 80% 的样本作为训练集样本，剩余的 20% 的样本作为验证集样本，取验证集最好的结果作为 Inception-CSA 模型的最终分类精度。预处理阶段使用 Python 库中的 Librosa 对每个声音样本提取 64 个梅尔滤波器组的梅尔频谱图。音频采样率为 22 050 Hz，帧移为 512，帧之间的重叠率为窗口大小的 1/4，4 s 长度的声音样本大致包含 174 帧。但数据集的所有音频样本的时长并非严格控制在 4 s，音频样本经过预处理获取到的梅尔频谱图的尺寸在时间维度上不统一，而网络模型的输入需要统一尺寸。将梅尔频谱图在时间维度上进行 0 填充，统一补齐至 192 帧，因此所有梅尔频谱图的尺寸均变为 64×192。0 填充的部分不会将噪声引入梅尔频谱图，因为其在卷积、池化、线性激活等操作的过程中，特征值一直保持为 0，不会在特征提取的过程中引入噪声。

1.5　试验条件与参数设置

本研究所有试验均在深度学习服务器上进行，服务器的试验条件为：CPU Intel i9，内存 32 GB，GPU GeForce RTX 3090 16 GB，操作系统 Windows 10 专业版，编辑语言 Python 3.7，深度学习框架 Pytorch 1.9。为了能使深度学习模型训练完全，试验进行 300 个 Epoch。试验中 Batch_size 对最终分类结果影响较小，为了尽可能地利用显存资源，将 Batch_size 设为 32，优化器选用 SGD 优化器。理想情况下学习率应该随着训练的进行不断衰减最终接近 0，模型的参数才能收敛到最佳效果，Pytorch 深度学习框架自带的库中 Exponential LR 学习率下降策略符合该要求，因此，本试验的学习率更新策略采用 Exponential LR：

$$new_{lr} = initial_{lr} \times \gamma^{epoch} \tag{6}$$

其中，$initial_{lr}$ 表示初始学习率，γ 表示学习率衰减参数，试验中经过多次调参优化二者分别设为 0.001 与 0.965。new_{lr} 表示每个 Epoch 的学习率。

损失函数选用分类任务中常用的交叉熵函数（cross entropy）：

$$Loss = \frac{1}{N}\sum_{i} -[\, y_i \cdot \log p_i + (1-y_i) \cdot \log(1-p_i)\,] \tag{7}$$

其中，Loss 表示损失值，y_i 表示第 i 个样本的标签，p_i 表示模型预测的结果，即第 i 类别的概率。

2　结果与分析

图 5 是基于本研究提出的 Inception-CSA 深度学习模型在自建数据集上的训练过程损失值变化和准确率变化，其中红线表示训练集样本，蓝线表示验证集样本。由图 5 可知，在前 150 个 Epoch，训练集和验证集损失值快速降低，准确率快速升高，并趋近于收敛。在后 150 个 Epoch，训练集损失值收敛接近 0，准确率收敛接近 100%，验证集损失值和准确率也逐渐收敛稳定。在整个训练过程中，训练集与验证集的损失值和准确率相差不多，未出现过拟合现象，表明建立的模型具有很好的鲁棒性。

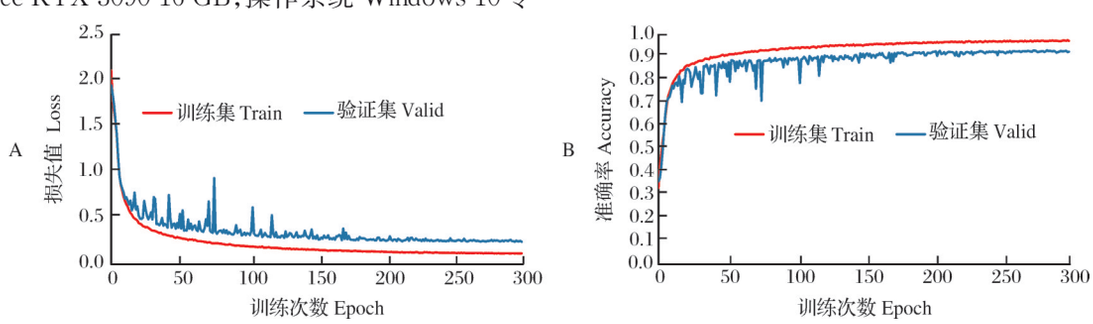


图 5　训练过程损失值（A）和准确率（B）变化

Fig.5　Loss value change（A）and accuracy rate change（B）during training

本研究使用混淆矩阵[19]对测试集的分类结果进行展示与分析（图 6）。该混淆矩阵横坐标代表真实值标签，纵坐标代表模型的预测值标签，位于对角线上的元素代表每一类分类正确的个数，其颜色越深则代表模型对该类的分类效果越好。验证集的混淆矩阵中，对角线的颜色深，表明本研究提出的分类方

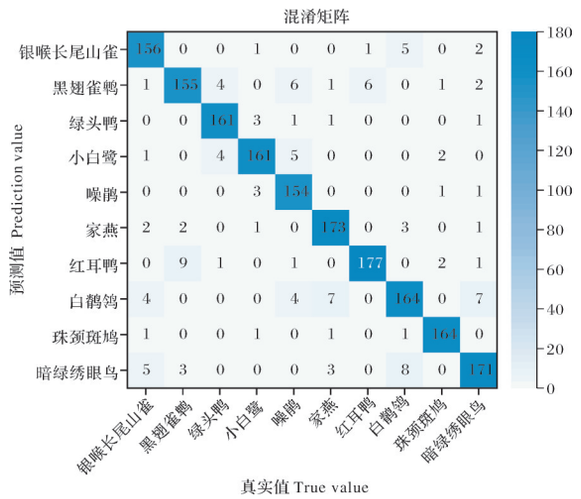


图6 最终验证集的混淆矩阵

Fig.6 Confusion matrix for final validation set

法具有高有效性和高精度性。

通过混淆矩阵计算出每个类别的精确率(precision)、召回率(recall)、F_1分数(F_1-score),如表1所示。由表1和表2可知,大部分类别的鸟鸣声分类各项指标超过90%,平均准确率为93.11%,表明对于鸟鸣声分类任务Inception-CSA模型具有加高的精度。但黑翅雀鹎与白鹡鸰的精确率仅有88.07%和88.17%,相比于其他类别偏低,可能是由于这2种鸟类的样本中掺杂了许多自然环境噪声,影响了分类任务的准确率,表明Inception-CSA模型在面对有较多噪声的音频样本的时候,难以准确地分类。

表1 混淆矩阵分析结果

Table 1 Confusion matrix analysis results

类别 Category	精确率/% Precision	召回率/% Recall	F_1分数/% F_1-score	数量 Number
银喉长尾山雀 *Aegithalos glaucogularis*	94.55	91.76	93.13	165
黑翅雀鹎 *Aegithina tiphia*	88.07	91.72	89.86	176
绿头鸭 *Anas platyrhynchos*	96.41	94.71	95.56	167
小白鹭 *Egretta garzetta*	93.06	94.71	93.87	173
噪鹃 *Eudynamys scolopaceus*	96.86	90.06	93.34	159
家燕 *Hirundo rustica*	95.05	93.01	94.01	182
红耳鸭 *Malacorhynchus membranaceus*	92.67	96.20	94.40	191
白鹡鸰 *Motacilla alba*	88.17	90.61	89.37	186
珠颈斑鸠 *Streptopelia chinensis*	97.62	96.47	97.04	168
暗绿绣眼鸟 *Zosterops japonicus*	90.00	91.94	90.96	190

不同数量的梅尔滤波器组会直接影响梅尔频谱图的大小,从而影响特征图的特征数量。梅尔滤波器组个数过少或过多,均会对分类精度产生影响。本研究将声音样本转换成不同个数梅尔滤波器组的梅尔频谱图,来考察分类精度最高的情况。当梅尔滤波器组数量分别为8、16、32、64、128时,对应准确率分别为90.32%、92.54%、92.83%、93.11%、93.06%。Inception-CSA模型对鸟鸣声的准确率最高为93.11%(表2),对应的梅尔滤波器组的数量为64个。

由表2可知,相比于一些经典分类网络模型,包括ResNet[20]、VGG[21]、AlexNet[22]、GoogleNet[17],本研究提出的Inception-CSA模型在拥有更少参数的同时,拥有更高的准确率。从基于不同深度的ResNet网络模型的试验结果可知,随着网络深度的增加,并不能给模型带来更好的分类效果,其原因可能是网络模型中的卷积核大小单一,导致模型不能学习到特征多样的鸟鸣声特征。

为验证基于Inception结构的模型相比于单种大小卷积核的卷积神经网络模型具有更好的效果,还构建相同规模的单种大小卷积核的卷积神经网络模型进行对比试验,结果表明基于Inception结构的网络模型分类效果要更好。在Inception结构中添加协调注意力后构成Inception-CA模型,该模型利用特征提取模块在关注多尺度局部高维特征的同时,还能利用全局关注权重来强化提取到的音频特征。试验结果显示,在Inception模块中加入协调注意力后,模型的分类精度提升了1.2个百分点,这表明协调注意力获取到的全局权重能对模型提取到的特征起到增强作用。Inception-CA模型与Inception-CSA模型的试验结果对比显示,sin函数使模型分类精度提升

表2 与其他分类网络模型的试验结果对比

Table 2 Comparison with experimental results of other classification network models

模型 Model	参数量 Parameters	准确率/% Accuracy
CNN	4.93×10^6	87.48
ResNet18	11.70×10^6	87.65
ResNet34	20.81×10^6	86.34
ResNet50	25.57×10^6	91.97
ResNet101	44.56×10^6	89.41
VGG16	138.30×10^6	88.50
AlexNet	61.11×10^6	86.37
GoogleNet	13.01×10^6	90.01
Inception-CA	1.58×10^6	91.27
Inception-CSA	1.57×10^6	93.11

1.9个百分点，表明sin函数在鸟鸣声特征提取过程中具有有效性。总之，Inception-CSA模型能在特征提取的过程中结合多尺度局部特征与全局关注权重，从而更容易学习与获取到鸟鸣声的特征，最终实现较高的分类准确率。

3　讨　论

本研究在原生Inception模块和协调注意力的基础上提出基于Inception-CSA模型的鸟鸣声分类方法。与基于单一感受野的卷积神经网络模型相比，Inception-CSA模型基于多感受野的卷积神经网络，分别在特征图中提取不同尺度的局部时频域特征，对于不同大小卷积核提取出来的特征图在特征图通道上进行拼接，同时利用改进后的协调注意力在特征图中获取到全局上的特征权重，然后将提取得到的特征图与特征权重进行矩阵乘法得到一个新的特征图，并用于之后的池化下采样与分类操作。这使得网络既能够感知特征图中不同尺度下的鸟鸣声时频域特征，又捕获了特征图的全局注意力权重，从而使网络获取丰富的鸟鸣声特征信息。其中，改进后的协调注意力（CSA）采用sin函数作为激活函数。sin函数的值域和周期性对相差较大的特征值进行约束，能简化鸟鸣声的特征差异。同时sin函数能保留经过傅里叶变换的鸟鸣声特征图中的正弦波特征。

本研究采集华南区域自然环境中常见的10种鸟类的鸣叫声，并构建出鸟鸣声分类数据集，提出的基于Inception-CSA模型的鸟鸣声分类方法在该数据集上的准确率为93.11%，和现有方法相比有较大的提升。并且在训练过程中，Inception-CSA模型收敛迅速，最终收敛至1个确定的值，表明该模型具有较强的鲁棒性。试验结果表明，即使同种鸟鸣声及不同种类鸟鸣声存在较大差异，模型依旧具有较高的分类精度，并且泛化性强。

本研究提出Inception-CSA神经网络模型用于鸟鸣声分类任务，并在自建鸟鸣声分类数据集上进行分类评测。相比于经典分类网络，本研究提出的模型参数量较少并且分类精度更高。在原生Inception的基础上利用协调注意力对卷积层进行改进，使模型既能够捕获多感受野提取的鸟鸣声特征，又能够获取鸟鸣声特征图全局注意力权重，从而增强鸟鸣声特征。在后续的研究中，我们会继续采集自然环境中的鸟鸣声数据以扩建鸟鸣声分类的数据集，致力于探究分类精度更高的网络模型。

参考文献 References

[1] 安文雨，涂婧林，侯东瑞，等. 国土空间生态修复与乡村振兴：共现与融合[J]. 华中农业大学学报，2022，41(3)：1-10. AN W Y，TU J Y，HOU D R，et al. Ecological restoration of territorial space and rural revitalization：co-occurrence and integration [J]. Journal of Huazhong Agricultural University，2022，41(3)：1-10 (in Chinese with English abstract).

[2] ANAND R，SHANTHI T，DINESH C，et al. AI based birds sound classification using convolutional neural networks[J/OL]. IOP conference series：earth and environmental science，2021，785(1)：012015[2022-09-19]. https://iopscience.iop.org/article/10.1088/1755-1315/785/1/012015/meta. DOI：10.1088/1755-1315/785/1/012015.

[3] BARDELI R，WOLFF D，KURTH F，et al. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring [J]. Pattern recognition letters，2010，31(12)：1524-1534.

[4] WIMMER J，TOWSEY M，ROE P，et al.Sampling environmental acoustic recordings to determine bird species richness[J].Ecological applications，2013，23(6)：1419-1428.

[5] 刘志华，陈文洁，陈爱斌. 基于自注意力机制时频谱同源特征融合的鸟鸣声分类[J]. 计算机应用，2022，42(4)：1260-1268. LIU Z H，CHEN W J，CHEN A B. Homologous spectrogram feature fusion with self-attention mechanism for bird sound classification [J]. Journal of computer applications，2022，42(4)：1260-1268(in Chinese with English abstract).

[6] BRIGGS F，LAKSHMINARAYANAN B，NEAL L，et al. Acoustic classification of multiple simultaneous bird species：a multi-instance multi-label approach[J].The journal of the Acoustical Society of America，2012，131(6)：4640-4650.

[7] QIAO Y，QIAN K，ZHAO Z.Learning higher representations from bioacoustics：a sequence-to-sequence deep learning approach for bird sound classification [C]//27th International Conference，ICONIP 2020，November 18-22，2020，Bangkok，Thailand.Cham：Springer，2020：130-138.

[8] ACEVEDO M A，CORRADA-BRAVO C J，CORRADA-BRAVO H，et al.Automated classification of bird and amphibian calls using machine learning：a comparison of methods[J].Ecological informatics，2009，4(4)：206-214.

[9] 魏静明，李应. 利用抗噪纹理特征的快速鸟鸣声识别[J]. 电子学报，2015，43(1)：185-190. WEI J M，LI Y.Rapid bird sound recognition using anti-noise texture features[J].Acta electronica sinica，2015，43(1)：185-190(in Chinese with English abstract).

[10] LEE C H，HSU S B，SHIH J L，et al.Continuous birdsong recognition using Gaussian mixture modeling of image shape features [J].IEEE transactions on multimedia，2012，15(2)：454-464.

[11] 张赛花，赵兆，许志勇，等. 基于Mel子带参数化特征的自动鸟鸣识别[J]. 计算机应用，2017，37(4)：1111-1115. ZHANG S H，ZHAO Z，XU Z Y，et al. Automatic bird vocalization identification based on Mel-subband parameterized feature[J]. Journal of computer applications，2017，37(4)：1111-1115(in Chinese with English abstract).

[12] JANČOVIČ P，KÖKÜER M，RUSSELL M. Bird species recognition from field recordings using HMM-based modelling of frequency tracks[C]//2014 IEEE International Conference on

Acoustics, Speech and Signal Processing, May 04-09, 2014, Florence, Italy. New York: IEEE, 2014: 8252-8256.

[13] HINTON G, DENG L, YU D, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups[J]. IEEE Signal processing magazine, 2012, 29(6): 82-97.

[14] ZHANG X, CHEN A, ZHOU G, et al. Spectrogram-frame linear network and continuous frame sequence for bird sound classification[J/OL]. Ecological informatics, 2019, 54: 101009[2022-09-19]. https://doi.org/10.1016/j.ecoinf.2019.101009.

[15] SPRENGEL E, JAGGI M, KILCHER Y, et al. Audio based bird species identification using deep learning techniques[C]// Conference and Labs of the Evaluation Forum (CLEF) 2016, September 5-8, 2016, Évora, Portugal. [S.l.]: LifeCLEF, 2016: 547-559.

[16] JOLY A, GOËAU H, GLOTIN H, et al. Lifeclef 2017 lab overview: multimedia species identification challenges[C]//International Conference of the Cross-Language Evaluation Forum for European Languages, Sept 11-14, 2017, Dublin, Ireland. Cham: Springer, 2017: 255-274.

[17] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 07-12, 2015, New York, USA. New York: IEEE, 2015: 1-9[2022-09-19].

[18] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design [DB/OL]. arXiv, 2021: 2103.02907 [2022-09-19]. https://doi.org/10.48550/arXiv.2103.02907.

[19] LIM M, LEE D, PARK H, et al. Convolutional neural network based audio event classification[J]. KSII transactions on internet and information systems (TIIS), 2018, 12(6): 2748-2760.

[20] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[DB/OL]. arXiv, 2015: 1409.1556[2022-09-19]. https://doi.org/10.48550/arXiv.1409.1556.

[21] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[DB/OL]. arXiv, 2015: 1512.03385 [2022-09-19]. https://doi. org/10.48550/arXiv.1512.03385.

[22] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

Inception-CSA deep learning model-based classification of bird sounds

LI Huaicheng, YANG Daowu, WEN Zhifang, WANG Ya'nan, CHEN Aibin

College of Computer and Information Engineering/Institute of Applied Artificial Intelligence, Central South University of Forestry and Technology, Changsha 410004, *China*

Abstract Bird sounds have diverse features, and most of the current convolutional neural network models based on a single receptive field are difficult to learn the diversity of bird sound features from audio containing complex background noise. In this article, we proposed a method of classifying bird sounds based on the Inception-CSA deep learning model, which consists of three steps including bird audio sample preprocessing, feature extraction, and classifier classification. First, the samples of bird sounds were preprocessed into Mel spectrum maps with the same size as the feature maps of bird sounds. Then the feature of bird sounds was extracted with the Inception-CSA model including the Inception module extracting the multi-scale local time-frequency domain features in the feature map of bird sounds and the CSA module obtaining the global attention weights of the feature map of bird sounds. The output of both was combined to obtain a stronger feature map. The feature maps were downsampled with the maximum pooling layer. Finally, the results of final classification were obtained with the fully connected layer. The calls of 10 wild bird species in the natural environment of south China were collected and the dataset was constructed to verify the effectiveness of the method. The results showed that the proposed method achieved 93.11% accuracy on the self-built dataset. The classification method based on the Inception-CSA model had higher accuracy with fewer model parameters compared with the classification methods based on other classical models.

Keywords convolutional neural network; classification of bird sound; deep learning; Mel spectrogram; Inception

(责任编辑:陆文昌)