DOI: 10.13382/j.jemi.B2407666

特征融合的密集连接卷积网络识别鸟鸣声

陈 晓[1,2] 颜 灏[1] 曾昭优[1]

(1.南京信息工程大学电子与信息工程学院 南京 210044;2.南京信息工程大学江苏省大气环境与装备技术协同创新中心 南京 210044)

摘 要:针对目前鸟鸣声识别的深度学习方法提取深层特征单一导致准确率不高的问题,提出一种改进密集连接卷积网络的鸟鸣声识别方法。从鸟鸣声信号中提取梅尔语谱图作为输入,在所有密集块的标准卷积层之后添加卷积块注意力模块,卷积块注意力模块通过学习训练集的特征表示,判断不同层次鸟鸣声特征信息的重要性和关联性,并按照通道维度和空间维度对其进行更深一步的加权融合,使网络更加关注鸟鸣声特征中重要的特征通道和空间位置,从而提高网络学习鸟鸣声特征的能力;在密集块的标准卷积层之后添加丢弃块算法,促使网络对于不同区域的特征进行更加均衡的学习,提高网络对于新鸟鸣声数据的适应能力,使网络能够更好地捕获数据中的共性特征;再利用 Transformer 编码器为网络建立一条深层特征提取分支,以提高对于鸟鸣声特征中全局信息和长距离依赖信息的捕捉能力。最后将两个分支提取的深层特征融合以提升深层特征的信息丰富度。该方法在 Xeno-Canto 数据集进行了 7 组实验。实验结果表明方法对鸟鸣声识别的平均准确率为 88.65%。相较于 EMSCNN(ensemble multi-scale convolutional neural network)方法高 10.83%,AlexNet 方法高 20.14%,VGGNet 方法高 16.3%,DenseNet 方法高 4.28%。实验证明了方法的有效性和先进性。提出的方法对鸟鸣声识别更准确,可用于实际鸟鸣声的识别。

关键词:声音识别;鸟声识别;密集连接卷积网络;特征融合;Transformer;深度学习

中图分类号:TN912.34 **文献标识码**:A **国家标准学科分类代码**:510.4040

Birdsong recognition based on improved DenseNet with feature fusion

Chen Xiao[1,2] Yan Hao[1] Zeng Zhaoyou[1]

(1. School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; 2. Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract: To address the issue of low accuracy caused by the single extraction of deep features in current bird sound recognition methods, this study proposed a DenseNet based bird sound recognition method with feature fusion. First, the Mel-spectrogram was extracted from bird sound signals as the network input. Then, DenseNet was used as the base network, and convolutional block attention module was integrated into the standard convolutional layer of all dense blocks dense blocks. The convolutional block attention module learns the feature representation of training set, determines the importance and correlation of different levels of bird song feature information, and further weights and fuses them according to channel and spatial dimensions, making the network pay more attention to the important feature channels and spatial positions in bird song features. Then, adding dropout block algorithm after the standard convolutional layer of dense blocks promotes the network to learn features from different regions in a more balanced manner, improves the network's adaptability to new bird song data, and enables the network to better capture common features in the data. Subsequently, a deep feature extraction branch using transformer encoder was established for DenseNet to enhance the network's ability to capture global information and long-distance dependencies in birdsong features. Finally, the deep features extracted by the two branches are fused to enrich the information content of the deep features. This method was tested in seven sets on the Xeno-Canto data set. Experimental results on the test data set show that the proposed method achieves an average accuracy of 88.65%, which is 10.83% higher than the

收稿日期:2024-07-10 Received Date: 2024-07-10

EMSCNN method, 20.14% higher than the AlexNet method, 16.3% higher than the VGGNet method, and 4.28% higher than DenseNet. The experiment proved the effectiveness and progressiveness of the proposed method. It outperforms other comparative deep learning methods in terms of recognition performance and effectiveness.

Keywords: acoustical recognition; birdsong recognition; DenseNet; feature fusion; transformer; deep learning

0 引 言

鸟类是生态系统不可或缺的一部分,广泛分布于森林、草原、湿地、沼泽、湖泊、河流、沙漠等地区。通过监测鸟类的种类、数量、迁徙和季节性分布变化,可以了解生态系统是否处于良好的状态,或者是否受到威胁和压力[1]。鸟鸣声识别在生物多样性保护、生态监测和和促进可持续发展方面具有重要意义[2]。

采用简单框架的鸟鸣声识别方法,如动态时间规整法[3]、支持向量机[4-5],能在低成本的硬件系统[6]如单片机[7]、嵌入式[8]等上实现,但在面对复杂环境时由于受硬件资源算力的限制,识别准确率不高,识别效果不佳。

由于深度学习算法[9]在多个领域中广泛应用[10]并表现出色[11],研究人员开始用深度学习方法处理鸟鸣声以提高识别能力。Sprengel 等[12]利用 5 个卷积层构建了适合鸟鸣声识别的卷积神经网络(convolutional neural network, CNN)模型,得到了较好的识别结果。Chandu 等[13]从声信号中提取语谱图输入预训练的 AlexNet(Alex network),获得了比 CNN 更高的准确率。Rajan 等[14]从鸟鸣声信号中提取梅尔语谱图输入 VGGNet(visual geometry group network),基于 Xeno-Canto 数据集的测试平均 F1 分数达到了 0.65。Liu 等[15]结合了双向长短期记忆网络(bidirectional long short-term memory, BiLSTM)[16]和密集连接卷积网络(densely connected convolutional networks, DenseNet),在 Birdsdata 数据集的识别准确率达到了 92.2%。Zhang 等[17]用 Transformer 编码器,在 Cornell Bird Challenge 数据集的准确率达到 93.18%。虽然上述基于深度学习的方法一定程度上提高了识别的准确率,但是在面对复杂背景和各种噪声时方法识别的准确率还有待提高。这主要是因为以上方法提取的深层特征较为单一,全局和局部特征的融合还有待进一步优化,算法的泛化能力也需要提高。

针对上述问题,本文提出改进 DenseNet 的鸟鸣声识别方法。首先从鸟鸣声信号中提取梅尔语谱图作为网络输入;然后在 DenseNet 的密集块中融合卷积块注意力模块(convolutional block attention module, CBAM)和丢弃块(drop block, DropBlock)算法,再利用 Transformer 编码器建立一条深层特征提取分支;最后将两个分支提取的深层特征融合。通过对比实验验证了本文方法的有效性和先进性。

本文创新点如下:1)在 DenseNet 的密集块中融合卷积块注意力模块,提升网络对于鸟鸣声特征中重要特征通道和空间位置的关注;2)在 DenseNet 的密集块中添加 DropBlock 算法,提高网络对于新数据的泛化能力;3)在 DenseNet 的密集块之后增加 Transformer 编码器分支,提高网络对于鸟鸣声特征中全局信息和长距离依赖信息的捕捉能力。

1 改进的鸟鸣声方法

本文方法首先计算鸟鸣声的梅尔语谱图,然后在 DenseNet[18]的基础上进行了以下 3 个方面改进,并用改进的网络识别鸟鸣声。1)在所有密集块的标准卷积层之后添加卷积块注意力模块,卷积块注意力模块通过学习训练集的特征表示,判断不同层次鸟鸣声特征信息的重要性和关联性,并按照通道维度和空间维度对其进行更深一步的加权融合,使网络更加关注鸟鸣声特征中重要的特征通道和空间位置,从而提高网络学习鸟鸣声特征的能力;2)在密集块的标准卷积层之后添加 DropBlock 算法,在训练过程中随机地屏蔽特征图的一部分区域,促使网络对于不同区域的特征进行更加均衡的学习,而不是依赖于少数特定区域的特征。这有助于提高网络对于新鸟鸣声数据的适应能力,使网络能够更好地捕获数据中的共性特征;3)在第 3 个密集块之后增加 Transformer 编码器分支,以提高网络对于鸟鸣声全局信息和长距离依赖信息的捕捉能力。分支首先将鸟鸣声特征的高度和宽度两个维度展开,再将展开后的特征输入 Transformer 编码器进行编码,得到鸟鸣声编码特征,然后在编码特征的序列长度维度上进行平均池化,得到特征向量,然后将线性层输出的鸟鸣声编码特征向量和 DenseNet 提取的深层特征拼接,得到融合特征。改进 DenseNet 整体结构如图 1 所示,密集块中的卷积组结构如图 2 所示。网络结构参数如表 1 所示。

1.1 梅尔语谱图

与用图像处理方法[19-20]获得图像的方式不同,梅尔语谱图从一维鸟鸣声信号中经过一系列变换得到展开的二维图像。提取梅尔语谱图需要 3 个步骤,分别是傅里叶变换、功率谱计算和梅尔滤波器组滤波。

将鸟鸣声信号变换到频域:

$$X(k)=\sum_{n=0}^{N-1}x(n)e^{-i2\pi k/N},0\leqslant k\leqslant N \tag{1}$$

式中:x 代表时域信号;N 代表快速傅里叶变换的长度;k 表示频率索引;X 表示信号的频谱。

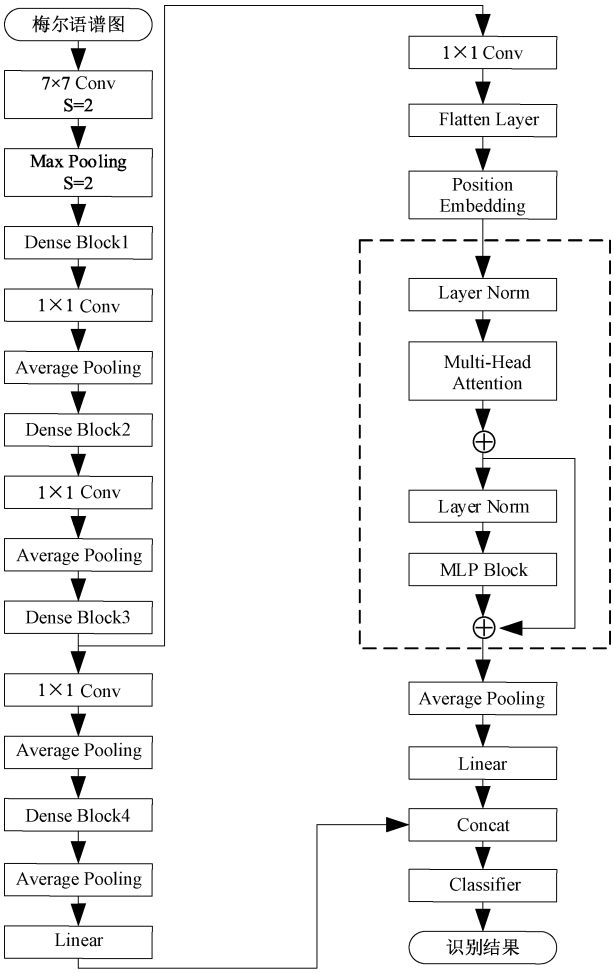

图 1　改进的 DenseNet

Fig. 1　Improved DenseNet network structure

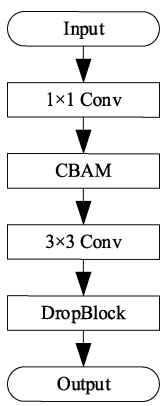

图 2　密集块中卷积组结构

Fig. 2　Convolutional group structure in dense block

计算鸟鸣声频域信号的功率谱:

$$P(k) = |X(k)|^2 \tag{2}$$

式中:P 表示信号的功率谱。

用梅尔滤波器组对功率谱滤波,得到梅尔语谱图:

$$E(m) = \sum_{k=1}^{N} P(k) \cdot H_m(k), 0 \leqslant m \leqslant M \tag{3}$$

表 1　DenseNet 网络参数

Table 1　DenseNet network parameters

层名	参数	输出尺寸
Convolution	7×7,S=2	112×112×64
Max Pooling	3×3,S=2	56×56×64
Dense Block1	$\begin{bmatrix}1\times1\\3\times3\end{bmatrix}\times6$	56×56×256
Transition Layer1	1×1 2×2,S=2	28×28×128
Dense Block2	$\begin{bmatrix}1\times1\\3\times3\end{bmatrix}\times12$	28×28×512
Transition Layer2	1×1 2×2,S=2	14×14×256
Dense Block3	$\begin{bmatrix}1\times1\\3\times3\end{bmatrix}\times24$	14×14×1 024
Transition Layer3	1×1 2×2,S=2	7×7×512
Dense Block4	$\begin{bmatrix}1\times1\\3\times3\end{bmatrix}\times16$	7×7×1 024
Global Average Pooling	7×7	1×1×1 024
Classification Layer	1 000 D	1×1×1 000

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)}, & f(m-1) \leqslant k \leqslant f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)}, & f(m) \leqslant k \leqslant f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{4}$$

$$f(m) = \left(\frac{N}{f_s}\right) F^{-1}\left(F(f_l) + m\frac{F(f_h) - F(f_l)}{M+1}\right) \tag{5}$$

式中:E 表示梅尔语谱图;f_s 表示采样频率;f_l 和 f_h 分别表示滤波器组的最低频率和最高频率;M 表示滤波器组内滤波器的数量。

1.2　卷积块注意力模块

CBAM[21] 通过对输入的梅尔语谱图进行通道和空间注意力的整合以提高网络的性能,如图 3 所示。

通道注意力模块关注输入特征的通道间关系。通道注意力图的计算公式如式(6)所示。

$$M_c(F) = \sigma[MLP(Avgpool(F)) + MLP(Maxpoll(F)] \tag{6}$$

式中:F 代表 CBAM 的输入特征;σ 代表 Sigmoid 激活函数。输入特征与通道注意力图相乘生成通道注意力特征 F'。通道注意力特征的计算公式如式(7)所示。

$$F' = M_c(F) \otimes F \tag{7}$$

式中:⊗表示逐元素乘法。

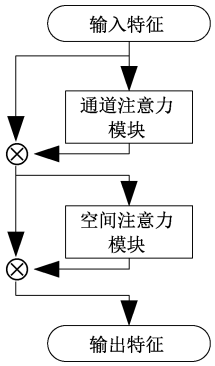

图3　卷积块注意力模块结构

Fig. 3　Structure of convolutional block attention module

空间注意力模块关注输入特征的空间位置信息。空间注意力图的计算公式如式(8)所示。

$$M_S(F') = \sigma(f^{7\times7}([Avgpool(F'), Maxpool(F')])) \tag{8}$$

输入特征与空间注意力图相乘生成空间注意力特征 F'',空间注意力特征的计算公式如式(9)所示。

$$F'' = M_s(F') \otimes F' \tag{9}$$

1.3　DropBlock 算法

DropBlock 算法为一种常用的正则化方法,通常用于全连接层,该算法通过阻止部分特征的信息传递,迫使网络学习更加鲁棒和泛化的特征表示,从而使得网络具有更好的泛化性能。DropBlock 算法是按照比例随机选择一个固定大小的块,将该块内的所有特征都丢弃,这种丢弃方式不残留信息,对于卷积更适用。

DropBlock 算法有两个参数,分别是 block_size 和 r。block_size 表示随机丢弃的块尺寸,当 block_size 为 1 时,DropBlock 算法转变为 Dropout 算法。r 控制需要丢弃的特征元素数量。r 的计算公式如式(10)所示。

$$r = \frac{(1 - keep\_prob) feat\_size^2}{block\_size^2 (feat\_size - block\_size + 1)^2} \tag{10}$$

式中:$keep\_prob$ 表示将特征元素保留的概率;$feat\_size$ 表示特征图的尺寸。

1.4　Transformer 编码器

Transformer[22] 编码器由归一化层、多头注意力(multi-head attention, MHA)和多层感知机组成,内部采用残差连接,结构如图4所示。

1)位置编码

由于 Transformer 编码器不包含任何显式的顺序信息,为了使编码器能够处理输入特征数据,需要添加位置编码来表示每个输入特征的位置信息。位置编码通常是与输入特征维度相同的向量,它被加到输入特征中以提供位置信息。在 Transformer 编码器中,位置编码分为正弦编码部分和余弦编码部分,正弦编码部分由一个正弦函数组成,用于编码序列中每个位置的奇数索引位置,余弦编码部分由一个余弦函数组成,用于编码序列中每个位置的偶数索引位置。位置编码的计算式如式(11)和(12)所示。

$$F_{PE}(p, 2i) = \sin\left(\frac{p}{10\ 000^{\frac{2i}{d}}}\right) \tag{11}$$

$$F_{PE}(p, 2i+1) = \cos\left(\frac{p}{10\ 000^{\frac{2i}{d}}}\right) \tag{12}$$

式中:p 表示输入特征中各个特征向量的位置;i 表示位置编码的维度;d 表示 Transformer 编码器的维度。

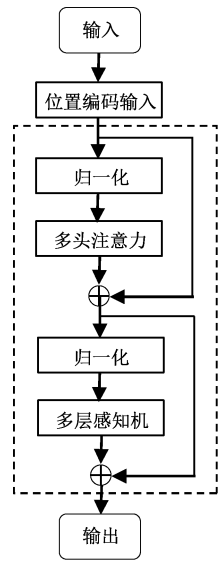

图4　Transformer 编码器结构

Fig. 4　Transformer encoder structure

2)多头注意力

MHA 是编码块的核心部分,属于自注意力的延伸。虽然自注意力能够捕捉输入特征中各个位置之间的相关性,但依旧存在诸多不足。Vaswani 等[23]在自注意力机制的基础上提出了 MHA,其结构如图5所示。

多头注意力通过多次相关性计算,得到多个不同的权重,这使网络能够在多个时空点上捕捉多个时空子空间的特征,最终融合多个特征,使网络能够更好地学习全局信息的长距离依赖关系。多头注意力的计算公式如式(13)和(14)所示。

$$MultiHead(Q, K, V) = Concat(head_1, \cdots, head_h) \boldsymbol{W}^O \tag{13}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{14}$$

式中:$\boldsymbol{W}^O$ 是多头注意力特征拼接后的线性变换矩阵,其作用是控制多头注意力特征的尺寸,使其与自注意力的尺寸相同。多头注意力相较于自注意力,前者将后者的

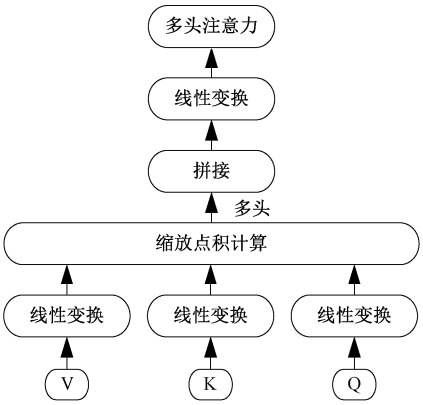

图 5　多头注意力计算过程

Fig. 5　Multi-head attention calculation process

一组变换矩阵计算扩充为多组并行计算,并且将单个值矩阵权重计算拓展为多个并行计算,这使得网络不仅能在不同时空位置获取上下文交互信息,还实现了高效计算,加快了网络的训练速度[24]。

3)多层感知机

编码块的多层感知机的结构包含 2 个线性层、1 个 GELU 激活函数层和 2 个丢弃(Dropout)层。第 1 个线性层将多头注意力特征的尺寸变换为原来的 4 倍,第 2 个线性层将特征尺寸还原,这种变换有助于网络捕捉不同特征之间的复杂关系。多层感知机的输出计算式如式(15)所示。

$$MLP(x) = GELU(x\boldsymbol{W}_1 + b_1)\boldsymbol{W}_2 + b_2 \tag{15}$$

式中:$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 表示两个线性层的权重矩阵;b_1 和 b_2 表示两个线性层的偏置项;GELU 表示 GELU 激活函数。GELU 激活函数的定义式如式(16)~(18)所示。

$$GELU(x) = x \cdot \Phi(x) \tag{16}$$

$$\Phi(x) = \frac{1}{2}\left(1 + f_{er}\left(\frac{x}{\sqrt{2}}\right)\right) \tag{17}$$

$$f_{er}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{e}^{-t^2}\mathrm{d}t \tag{18}$$

2　实验结果与分析

2.1　实验环境与参数设置

为了测试改进方法有效性、可行性和评价方法的性能,在标准数据集上进行了实验。实验是在 Ubuntu 操作系统下进行,GPU 型号是 NVIDIA RTX 2080Ti,网络模型采用 Tensorflow 框架搭建,编程语言使用 Python。网络训练使用 Adam 优化器更新权重,批的尺寸设置为 32,初始学习率为 0.001,循环为 100。采用 Xeno-Canto 数据集,可以先对鸟鸣声用去噪算法[25]去除噪声[26]、分离出有用的信号[27]。实验环境设置如表 2 所示。

表 2　实验环境

Table 2　Experimental environment

实验环境	描述
操作系统	Ubuntu
硬件设备	NVIDIA RTX 2080Ti
框架	TensorFlow
编程语言	Python
优化器	Adam
批尺寸	32
初始学习率	0.001
数据集	Xeno-Canto 数据集
样本组划分	5 个样本组:1、2、3、4、5

为了从多方面分析改进的方法,设计了 7 个实验进行验证测试,如表 3 所示。在 7 个实验中网络输入特征均为梅尔语谱图。用识别的准确率来评估方法的识别效果。

表 3　7 种实验对比

Table 3　Comparison of 7 experiments

实验编号	实验名称	描述
实验 1	有效性实验	验证方法的有效性
实验 2	消融实验	验证各模块的重要性
实验 3	不同 CBAM 方案的对比实验	比较不同 CBAM 的效果
实验 4	丢弃块尺寸的对比实验	比较不同丢弃尺寸的效果
实验 5	分支位置的对比实验	比较不同分支位置的效果
实验 6	不同融合方案的对比实验	比较不同融合方案的效果
实验 7	对比实验	与其他方法进行对比

2.2　结果分析

1)有效性实验

测试准确率如表 4 所示,各类鸟鸣声平均 F1 分数如表 5 所示。由表 4 可知,当以 1、2、3 和 4 号样本组为训练集,5 号样本组为测试集时,准确率为 89.02%,是 5 次测试中的最高值;当以 1、2、4 和 5 号样本组作为训练集,3 号样本组作为测试集时,准确率为 87.54%,是 5 次测试中的最低值。由表 5 可知,46 种鸟鸣声的平均 F1 分数最高值为 0.953 1,最低值为 0.798 8,这说明不同种类鸟鸣声的识别难度有一定的差异,不同种类鸟的鸣声提前的特征有很大差别。实验中 5 次测试的平均准确率为 88.65%,平均 F1 分数为 0.886,平均分数高于 0.9 的鸟鸣声有 19 种,这证明本方法识别鸟鸣声较有效。

表 4　方法有效性实验的准确率

Table 4　Accuracy of experiment

训练集/测试集	准确率/%
1、2、3、4/5	89.02
1、2、3、5/4	89.13
1、2、4、5/3	87.54
1、3、4、5/2	89.42
2、3、4、5/1	88.14
平均值	88.65

表5　各类鸟鸣声平均F1分数

Table 5　Average F1-score of various birdsongs

种类	F1分数	种类	F1分数
西鹌鹑	0.851 6	暗绿柳莺	0.856 3
水蒲苇莺	0.940 8	林百灵	0.873 5
普通海番鸭	0.915 5	大杜鹃	0.953 1
红隼	0.841 8	云雀	0.875 4
红交嘴雀	0.900 3	欧亚鸲	0.865 4
金黄鹂	0.893 1	鹪鹩	0.890 1
蚁鵙	0.878 0	灰斑鸠	0.923 1
欧歌鸫	0.868 3	黑啄木鸟	0.904 5
白颊黑雁	0.904 5	大苇莺	0.899 2
黄鸦	0.879 4	家燕	0.919 6
黄道眉鸦	0.902 9	寒鸦	0.798 8
黍鸦	0.816 8	夜鹭	0.914 5
芦鸦	0.910 1	斯氏夜鹱	0.866 0
大斑啄木鸟	0.921 1	家麻雀	0.904 0
普通翠鸟	0.941 2	紫翅椋鸟	0.865 3
欧夜鹰	0.845 9	冠小嘴乌鸦	0.913 3
白颊黑雁	0.905 3	欧亚喜鹊	0.846 4
斑尾林鸽	0.828 0	林柳莺	0.889 5
赤颈鸭	0.853 4	蓝喉歌鸲	0.854 6
绿翅鸭	0.878 5	鬼鸮	0.885 4
纵纹腹小鸮	0.873 7	赭红尾鸲	0.924 8
苍鹭	0.858 5	疣鼻天鹅	0.931 3
鹊鸲	0.915 4	红领绿鹦鹉	0.893 8

2)消融实验

为了CBAM、DropBlock算法和Transformer编码器分支的各自的作用和有效性,将DenseNet作为原方法进行消融实验,各种改进方法的结构如表6所示,表中Transformer表示Transformer编码器分支,DB表示DropBlock,实验结果如表7所示。根据实验结果可知,首先DenseNet结合卷积块注意力模块使得识别准确率提高了1.25%,这是由于卷积块注意力模块根据重要性和关联性对不同层次特征进行了融合,使得网络提取的特征信息更丰富。在DenseNet中添加DropBlock,准确率提高了0.61%,这证明DropBlock可以促使网络更加高效地学习鸟鸣声特征,提升网络的泛化性能。在DenseNet中增加Transformer编码器分支,准确率提高了2.42%,这证明Transformer编码器捕捉全局信息和长距离依赖信息的能力适用于鸟鸣声分类识别。消融实验证明了CBAM、DropBlock算法和Transformer编码器分支对提高准确率都有贡献。

表6　各种改进方法的结构

Table 6　Structure of various improvement methods

方法	DenseNet	CBAM	DB	Transformer
原方法	√	×	×	×
改进方法1	√	√	×	×
改进方法2	√	√	√	×
本文方法	√	√	√	√

表7　消融实验结果

Table 7　Ablation experiment results

方法	准确率/%
原方法	84.37
改进方法1	85.62
改进方法2	86.23
本文方法	88.65

3)不同CBAM嵌入方案的对比实验

为了验证CBAM嵌入的效果,设计了包括本文使用的方法在内总计10个嵌入方案,如表8所示。使用这10个方案进行对比试验,实验结果如表9所示。

表8　10种CBAM嵌入方案

Table 8　10 CBAM embedding schemes

嵌入方案编号	嵌入位置描述
方案1	网络的第1个卷积层后嵌入
方案2	网络的最大池化层后嵌入
方案3	仅在第1个密集块中每个1×1卷积层后嵌入
方案4	仅在第1和第2个密集块中每个1×1卷积层后嵌入
方案5	第1、第2和第3个密集块中每个1×1卷积层后嵌入
方案6	4个密集块中每个3×3卷积层后,特征拼接前嵌入
方案6	4个密集块中每次特征拼接后嵌入
方案8	每个过渡层的1×1卷积后嵌入
方案9	每个过渡层的平均池化层后嵌入
本文方案	4个密集块中每个1×1卷积层后嵌入

表9　不同CBAM嵌入方案的对比实验结果

Table 9　Comparative experimental results of different CBAM embedding schemes

CBAM嵌入方案	准确率/%
无嵌入	87.72
嵌入方案1	85.48
嵌入方案2	82.32
嵌入方案3	88.05
嵌入方案4	88.18
嵌入方案5	88.53
嵌入方案6	87.91
嵌入方案7	87.64
嵌入方案8	88.16
嵌入方案9	85.48
本文方案	88.65

由表8和9结果可知,相较于其他位置,本文所用的将CBAM嵌入所有卷积组的1×1卷积层的方案更有效,这是因为CBAM在空间维度和通道维度上对拼接特征加权融合,使得网络更准确地捕捉到鸟鸣声特征之间的复杂关系。

4)不同丢弃尺寸的对比实验

丢弃尺寸是DropBlock算法的重要参数,对Dropblock算法的性能至关重要。常用的丢弃尺寸包括

1×1、3×3、5×5 和 7×7。本文实验将 4 种丢弃尺寸的识别结果进行对比,实验结果如表 10 所示。

表 10　不同丢弃尺寸的对比实验结果

Table 10　Comparative results of different drop sizes

丢弃块尺寸	准确率/%
1×1	88.16
3×3	88.42
5×5	88.47
7×7	88.65

由实验结果可知,随着丢弃尺寸的增大,改进方法的准确率不断提高。在丢弃尺寸为 7×7 时,改进方法的识别准确率达到最高。因此丢弃尺寸设置为 7×7 最合适。原因在于更大的丢弃尺寸使得丢弃块中心元素的信息被更彻底地丢弃,这在更大程度上促使网络在不同的局部区域学习特征。

5)不同 Transformer 分支位置的对比实验

为了验证本文选择的 Transformer 编码器分支位置效果,分别在改进网络的 5 个位置建立 Transformer 编码器分支,并在其他条件相同的情况下进行对比试验,5 个分支位置如图 6 所示,实验结果如表 11 所示。

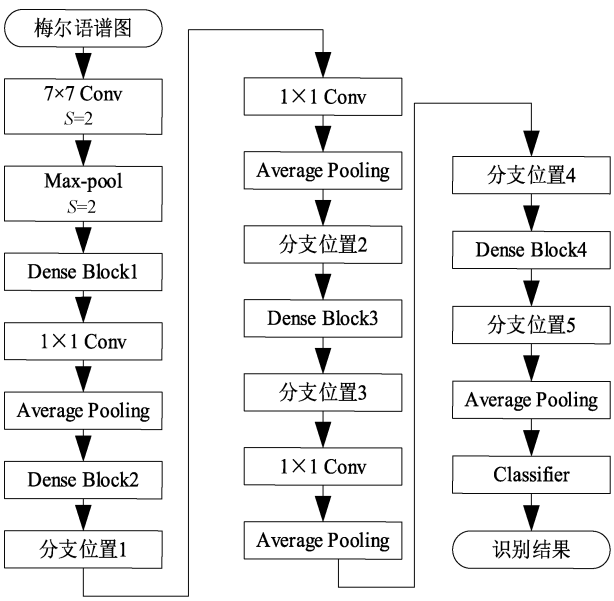

图 6　5 个分支位置

Fig. 6　Five branch locations

由实验结果可看出,在分支位置 1 建立 Transformer 编码器分支,虽然使得网络的准确率有所上升,但由于位置较浅,所提取的鸟鸣声特征较为简单,最终导致准确率比改进方法的低。分支位置 2 和 4 之前是一个平均池化层,平均池化会丢失原始鸟鸣声特征图的空间结构,无法保留鸟鸣声特征图的空间位置信息,这导致在分支位置 2 和 4 建立 Transformer 编码器分支的准确率相较于不建立分支低。分支位置 5 接近于输出层,与输出层之间仅存在一个平均池化层,在此位置建立 Transformer 编码器分支会导致信息传递不通畅,最终使得准确率相较于不建立分支大幅降低。改进方法的位置位于第 3 个密集块之后,此位置提取的鸟鸣声特征相较于分支位置 1 更深,相较于分支位置 5 更浅,这使得此位置的鸟鸣声特征不至于过于简单或者过于抽象,在这个位置建立 Transformer 编码器分支,所取得的识别准确率最高。这说明 Transformer 编码器分支的位置不能过浅或者过深,过浅的位置提取的特征不够丰富,过深的位置建立分支会影响鸟鸣声特征信息的传递,在 5 个分支位置中,改进方法的分支位置是最佳位置。

表 11　不同分支位置的对比实验结果

Table 11　Comparative experimental results of different branch positions

Transformer 编码器分支位置	准确率/%
无分支	85.89
分支位置 1	86.93
分支位置 2	83.25
分支位置 3(改进方法的分支位置)	88.65
分支位置 4	82.41
分支位置 5	78.63

6)不同特征融合的对比实验

本文实验通过两条分支提取鸟鸣声特征,然后在顶层将两路特征通过拼接的方案融合。为了验证本文方法的效果,另外设计了 4 种特征融合方案进行对比。设 DenseNet 分支提取的特征为 $\boldsymbol{a}$,Transformer 编码器分支提取的特征为 $\boldsymbol{b}$,其中 $\boldsymbol{a}$ 为$[a_1,a_2,a_3,\cdots,a_n]$,$\boldsymbol{b}$ 为$[b_1,b_2,b_3,\cdots,b_n]$,5 种特征融合方案如图 7 所示。将 5 种方案置于相同条件下进行实验测试,实验结果如图 8 所示。

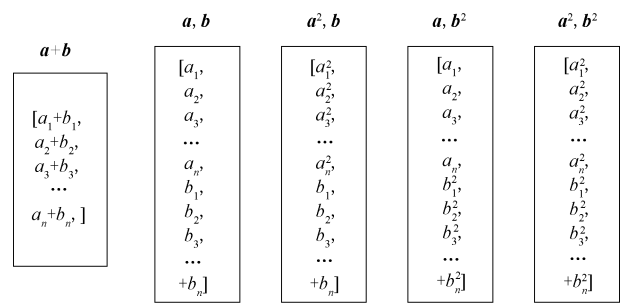

图 7　5 种特征融合方案

Fig. 7　Five feature fusion schemes

从实验结果可知,直接拼接方案的识别准确率最高,直接相加方案所的准确率仅比直接拼接方案低,两种单路平方后相加的方案效果不佳,先平方后拼接的方案所达到的识别准确率最低。上述实验结果证明本文方法所用的直接拼接特征融合方案最有效。

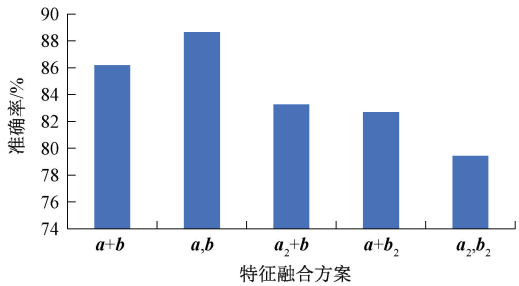

图 8　不同特征融合方案的对比实验结果

Fig. 8　Comparative experimental results of different feature fusion schemes

7) 与其他方法的对比实验

为了进一步验证本文方法的先进性,将本文方法与其他最新的4种深度学习方法在Xeno-Canto数据集进行了对比实验,实验结果如表12所示。从实验结果可知,本文方法的识别准确率在5种方法中最高,相较于EMSCNN(ensemble multi-scale convolutional neural network)[28]高10.83%,相较于AlexNet[13]的方法高20.14%,相较于VGGNet[14]高16.3%,相较于DenseNet高4.28%。可见本文方法相较于对比的4种方法识别效果更好。本文方法的参数量相对较少,比DenseNet少了8.3%。参数量少表示网络复杂度低,训练时间短。测试中5种方法的推理时间均在亚秒级,在实际使用中没有区别。

表12　与其他方法的对比实验结果

Table 12　Comparative experimental results with other methods

方法	准确率/%	参数量/($\times10^6$)
EMSCNN	77.82	13.37
AlexNet	68.51	56.97
VGGNet	72.35	134.36
DenseNet	84.37	7.98
本文方法	88.65	7.32

3　结　论

为了提高鸟鸣声识别的准确率,在密集连接卷积网络的基础上,通过添加卷积块注意力模块增强了网络对重要特征通道和空间位置的感知能力;通过在密集块的卷积之后添加丢弃块算法提高了网络的泛化能力;通过在密集块之后增加Transformer编码器分支提高了网络中特征全局信息和长距离依赖信息的捕捉能力。本文方法对鸟鸣声识别更准确,可应用于生态保护和生物多样性研究、环境监测、机场等区域驱鸟、电力行业输电线防鸟害等众多实际应用领域。下一步工作可以发展和改进轻量级的深度学习网络用于鸟声识别。

参考文献

[1] 陈晓, 曾昭优. 基于特征融合和B-SVM的鸟鸣声识别算法[J]. 声学技术, 2024, 43(1): 119-126.
CHEN X, ZENG ZH Y. Bird sound recognition algorithm based on feature fusion and B-SVM [J]. Technical Acoustics, 2024, 43(1): 119-126.

[2] 刘钊,张宇琛,胡海龙. 随机森林和大规模声学特征的噪声环境鸟声识别仿真[J]. 系统仿真技术, 2017, 13(4): 359-362.
LIU ZH, ZHANG Y CH, HU H L. Simulation of bird sound recognition in noise environment based on random forest and large-scale acoustic features [J]. System Simulation Technology, 2017,13(4): 359-362.

[3] TAN L, ALWAN A, KOSSAN G, et al. Dynamic time warping and sparse representation classification for birdsong phrase classification using limited training data [J]. The Journal of the Acoustical Society of America, 2015, 137(3): 1069-1080.

[4] 陈晓, 曾昭优. 基于声音特征优化和改进支持向量机的鸟声识别[J]. 测控技术, 2024, 43(6): 21-25.
CHEN X, ZENG ZH Y. Bird sound recognition based on optimized sound feature and improved SVM [J]. Measurement and Control Technology, 2024, 43(6): 21-25.

[5] 韩鹏飞, 陈晓. 基于MFCC-IMFCC和GA-SVM的鸟声识别[J]. 计算机系统应用, 2022, 31(11): 393-399.
HAN P F, CHEN X. Bird sound recognition based on MFCC-IMFCC and GA-SVM [J]. Computer Systems & Applications, 2022, 31(11): 393-399.

[6] 赵苏徽, 陈晓. 基于树莓派和云平台的智能灌溉系统[J]. 计算机系统应用, 2022, 31(4): 123-129.
ZHAO S W, CHEN X. Intelligent irrigation system based on raspberry Pi and cloud platform [J]. Computer Systems & Applications, 2022, 31(4): 123-129.

[7] 陈晓, 张凯. 嵌入式自动气象站故障检测系统研究[J]. 电子测量技术, 2021, 44(23): 158-164.
CHEN X, ZHANG K. Fault detector for automatic weather stations based on embedded system [J]. Electronic Measurement Technology, 2021, 44(23): 158-164.

[8] 荣百川, 陈晓. 基于蓝牙的智能防摔监测系统设计[J]. 激光杂志, 2019, 40(7): 32-34.
RONG B CH, CHEN X. Design of intelligent anti-fall

monitoring system based on bluetooth [J]. Laser Journal, 2019, 40(7): 32-34.

[9] 陈晓, 夏颖. 基于改进 MobileViT 网络的番茄叶片病害识别[J]. 电子测量技术, 2023, 46(14):188-196.
CHEN X, XIA Y. Improved MobileViT network for tomato leaf disease identification [J]. Electronic Measurement Technology, 2023, 46(14): 188-196.

[10] 陈晓, 杨瑶. 融合注意力机制的 BiLSTM 网络实现无创血压测量[J]. 电子测量技术, 2022, 45(23): 59-65.
CHEN X, YANG Y. Non-invasive blood pressure measurement based on BiLSTM network with attention mechanism [J]. Electronic Measurement Technology, 2022, 45(23): 59-65.

[11] 孙超文, 陈晓. 基于多尺度特征融合反投影网络的图像超分辨率重建[J]. 自动化学报, 2021, 47(7): 1689-1700.
SUN CH W, CHEN X. Multiscale feature fusion back-projection network for image super-resolution [J]. Acta Automatica Sinica, 2021, 47(7): 1689-1700.

[12] SPRENGEL E, JAGGI M, KILCHER Y, et al. Audio based bird species identification using deep learning techniques [J]. LifeCLEF, 2016: 547-559.

[13] CHANDU B, MUNIKOTI A, MURTHY K S, et al. Automated bird species identification using audio signal processing and neural networks[C]. 2020 International Conference on Artificial Intelligence and Signal Processing. IEEE, 2020: 1-5.

[14] RAJAN R, NOUMIDA A. Multi-label bird species classification using transfer learning [C]. 2021 International Conference on Communication, Control and Information Sciences. IEEE, 2021, 1: 1-5.

[15] LIU H, LIU C, ZHAO T, et al. Bird song classification based on improved Bi-LSTM-DenseNet network[C]. 4th International Conference on Robotics, Control and Automation Engineering. IEEE, 2021: 152-155.

[16] 任晓晔,陈晓,郭妍. 基于 Fluent 和 LSTM 神经网络的超声波测风仪阴影效应补偿研究[J]. 计算机应用与软件, 2019, 36(7): 89-98.
REN X H, CHEN X, GUO Y. Shadow effect compensation of ultrasonic wind measurer [J]. Computer Applications and Software. 2019, 36(7): 89-98.

[17] ZHANG S, GAO Y, CAI J, et al. A novel bird sound recognition method based on multifeature fusion and a transformer encoder [J]. Sensors, 2023, 23(19): 8099.

[18] HUANG G, LIU Z, VAN D M, et al. Densely connected convolutional networks [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4700-4708.

[19] 陈晓, 戴杰. 基于预测兰姆波参考信号的缺陷概率成像[J]. 电子学报, 2024, 52(9): 3262-3271.
CHEN X, DAI J. Probability imaging for defects using predicted Lamb wave reference signal [J]. ACTA Electronic SINICA, 2024, 52(9): 3262-3271.

[20] 陈晓, 戴杰. 基于相同传播距离路径的兰姆波无基准损伤概率成像[J]. 电子测量与仪器学报, 2023, 37(8): 94-104.
CHEN X, DAI J. Lamb wave baseline-free damage probability imaging based on same propagation distance path [J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(8): 94-104.

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

[22] 杨傲雷, 周应宏, 杨帮华, 等. 基于 Transformer 的三维人体姿态估计及其动作达成度评估[J]. 仪器仪表学报, 2024, 45(4): 136-144.
YANG AO L, ZHOU Y H, YANG B H, et al. Three dimensional human pose estimation and motion achievement evaluation based on transformer [J]. Chinese Journal of Scientific Instrument, 2024, 45(4): 136-144.

[23] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in Neural Information Processing Systems, 2017, 30: 6000-6010.

[24] CHEN H, JIANG D, SAHLI H. Transformer encoder with multi-modal multi-head attention for continuous affect recognition[J]. IEEE Transactions on Multimedia, 2020, 23: 4171-4183.

[25] 陈晓, 汪陈龙. 基于赛利斯模型和分数阶微分的兰姆波信号消噪[J]. 物理学报, 2014, 63(18): 282-290.
CHEN X, WANG CH L. Noise suppression for Lamb wave signals by Tsallis mode and fractional-order differential [J]. Acta Physica Sinica, 2014, 63(18): 282-290.

[26] 徐畅, 陈晓, 季仟亿. 基于稀疏编码的 Shearlet 域图像去噪[J]. 激光杂志, 2017, 38(10): 96-100.
XU CH, CHEN X, JI Q Y. Shearlet domain image denoising via sparse coding [J]. Laser Journal, 2017, 38(10): 96-100.

[27] 陈晓，倪龙. 用分数阶微分实现时频重叠多模式兰姆波的模式分离[J]. 声学学报，2020，45(2)：205-214.
CHEN X，NI L. Mode separation for multimode Lamb waves overlapped in time and frequency domains by using fractional differential [J]. Acta Acustica，2020，45(2)：205-214.

[28] LIU J，ZHANG Y，LYU D，et al. Birdsong classification based on ensemble multi-scale convolutional neural network [J]. Scientific Reports，2022，12(1)：8636.

作者简介

陈晓(通信作者)，南京信息工程大学教授，主要研究方向为现代电子系统设计、声信号与信息处理、图像处理、成像等。
E-mail：chenxiao@nuist.edu.cn

Chen Xiao (Corresponding author) is now a professor in Nanjing University of Information Science and Technology. His main research interests include modern electronic system design，acoustical signal and information processing，image processing and imaging.

颜灏，南京信息工程大学硕士研究生，主要研究方向为风电功率预测。

Yan Hao is now a M. Sc. candidate at Nanjing University of Information Science and Technology. His research interest includes wind power prediction.

曾昭优，南京信息工程大学硕士研究生，主要研究方向为信号处理。

Zeng Zhaoyou is a now a M. Sc. candidate at Nanjing University of Information Science and Technology. His main research interest includes signal processing.