通过主要颜色添加剂改善鸟类分类

Ezhini Rasendiran R[1], *Chandresh Kumar Maurya*[2]

[1]Department of Metallurgical Engineering and Materials Science, Indian Institute of Technology Indore, India
[2]Department of Computer Science & Engineering, Indian Institute of Technology Indore, India
mems210005019@alum.iiti.ac.in, chandresh@iiti.ac.in

Abstract

我们解决了使用鸟类鸣唱录音进行鸟类物种分类的问题，这是一个由于环境噪声、重叠的发声和缺失标签而具有挑战性的任务。现有模型在处理低信噪比或多物种录音时表现不佳。我们假设可以通过可视化鸟类的音调模式、速度和重复（统称为动机）来进行分类。虽然应用于频谱图图像的深度学习模型有帮助，但跨物种相似的动机会引起混淆。为了解决这一问题，我们将频率信息嵌入频谱图中，使用原色添加剂。这增强了物种的区分性，提高了分类的准确性。我们的实验表明，所提出的方法在不使用彩色化的模型上取得了统计上显著的提升，并超越了 BirdCLEF 2024 的获胜者，使 F1 提高了 7.3

音频分类是基于音频记录的声学特征将其分类到预定义类别中的过程。这种技术在生物多样性保护工作中被越来越多地使用，尤其是在野生动物的监测方面。通过使用先进的机器学习模型和音频处理技术，研究人员可以自动识别鸟鸣声、昆虫声及其他野生动物的叫声。这种方法对于追踪物种在其自然栖息地的存在、丰度和行为至关重要。传统的生物多样性监测方法通常涉及人工观察和数据收集，这既耗时又费力，并且容易出现人为错误。采用音频分类能够更加高效、准确地进行监测，为生态系统健康状况提供实时见解。例如，自动化系统可以分析部署在偏远地区的生物声学被动声学监测（PAM）机器中获得的大型数据集，使得在更长时间内监测生物多样性成为可能。通过自动化识别物种的声音，帮助研究人员做出明智的决策来保护生态系统，并抗击诸如栖息地丧失和气候变化的威胁。这一技术应用是实现全球可持续发展目标的一步，通过确保我们星球丰富的生物多样性得以保留。

深度学习的突破通过支持从原始数据中提取深层抽象特征革命性地改变了生物声学音频分类 [1]；然而，物种声音重叠、多样化的声音动机以及物种叫声之间的惊人相似性仍然使得稳健的分类变得复杂。

诸如 MixIT 方法 [2] 这样的无监督源分离技术已被用于在复杂的声景中分离重叠的鸟类鸣叫。通过隔离单独的叫声并实现声学特征的更清晰提取，这一方法提高了分类性能。它有效地减少了不同类别声音图案的干扰，从而导致更精确的物种识别 [3]。然而，过度分离在某些情况下可能会删除重要的上下文线索，降低对最显著物种的检测概率。在这些情况下，过度分离也可能将较长的鸣叫分割成孤立的音符，这些音符类似于其他物种的叫声，导致误分类。

因此，我们引入了一种新颖的特征工程方法，将频率信息直接嵌入输入的梅尔频谱图中。这一增强使模型能够捕捉到不同物种中相似声乐模式中的频率变化。通过强调这些潜在的差异，模型能够更好地学习和区分物种的叫声。我们的实验表明，这一策略显著提高了分类的鲁棒性。统计分析证实，与未进行此增强的模型相比，我们设计的特征带来了显著的性能提升。我们的主要贡献包括：

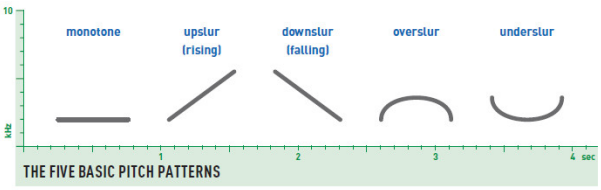


Figure 1: 五种基本音高模式。图取自 *[4]*，已获许可。

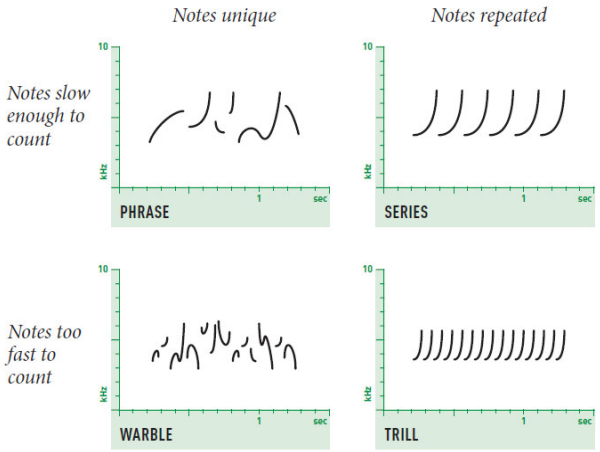


Figure 2: 重复和速度分类：短语、系列、颤音和颤鸣。图取自 *[4]*，经许可使用

- 提出一种新的想法，将频率信息嵌入到梅尔频谱图中，以解决灰度梅尔频谱图中的相似主题问题。
- 我们通过实证研究表明，我们提出的方法在处理相似的模式图案方面是有效的，并且优于 BirdCLEF 2024 的获胜模型。

1. 背景和问题陈述

1.1. 鸟类声音可视化

鸟类可以通过视觉化它们的声音来识别 [4]。关键方面包括音高模式、速度、重复、停顿和音质。频谱图符号提供了一种简化的鸟类声音艺术表现，强调模式而非细节，使其对人类有用但对计算机无用。对于实验，使用真实的频谱图对鸟类进行分类。与音乐中确切的音高不同，鸟类声音识别更注重随时间变化的音高变化。这种方法通过捕捉独特的听觉模式来增强物种识别，同时利用深度学习进行准确分类。所有鸟类的声音可以分为五个子类别（或其组合）如图 1 所示。

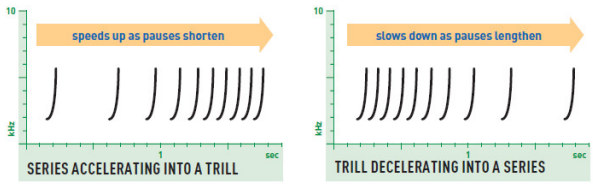


Figure 3: 级数加速成颤音，颤音减速成级数。图由 *[4]* 授权摘录

单调 声音保持恒定的音高，并在声谱图上显示为水平线。上扬的声音 声音音高增加，显示出向上的倾斜。 **下滑音** 声音音高降低，显示出向下的倾斜。过度连奏 声音音高先升高后降低，最高点出现在中间。 **下划线不清晰** 声音音高先降低后升高，最低点出现在中间。

重复（又称为动机）和速度分别与鸟类多次唱相同音符的事实及其发生的速度有关。整体来说，重复和速度有四种基本模式：短语、系列、颤音和音颤。短语和系列是较慢的声音，单个音符清晰可辨。短语包含不重复的独特音符，而系列由单个音符多次重复组成。颤音和音颤是短语和系列的快速版本，音符的出现速度快到无法计数（通常比每秒八个音符还快）。这些动机复杂地组合在一起形成鸟类的歌声，音高和速度各异。时间变化包括加速和减速（如图 3 所示）。这种动态模式突显了鸟类声学的复杂性，使鸟鸣在物种鉴定和行为研究中具有不可替代的价值。鸟类物种音频识别最终归结为识别录音中的这些独特动机。

具有多个实例 x_j （录音被切割为固定大小的窗口，如 §2 中所述），以及录音标记的弱标签 $Y_i \in \{0,1\} \in \mathcal{Y}$ 的数据集 $\{X_i, Y_i\}_{i=1}^{n}$ 。若任何一个实例 x_i 为正则为 $Y_i = 1$ ，若所有实例为负则为 $Y_i = 0$ 。我们的目标是建立一个多类多标签分类器 $f : \mathcal{X} \to \mathcal{Y}$ 。

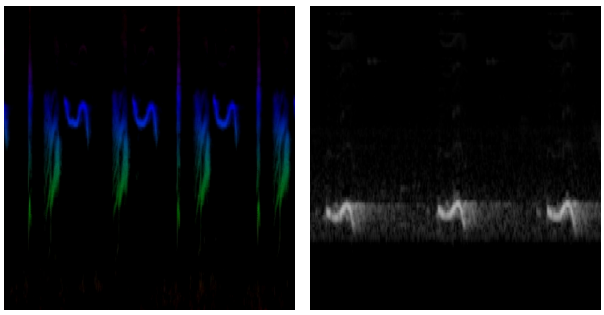
2. 方法论

这部分介绍我们解决鸟类声音的音频分类的方法。我们假设仅使用梅尔频谱图作为音频特征并将其输入深度学习模型是不够的。原因是灰度梅尔频谱图（作为单一通道）丢失了频率信息（当作为卷积神经网络（CNN）的输入图像使用时），而频率信息是声音特征识别的重要组成部分。使用灰度梅尔频谱图来分类不同鸟类声音的第二个问题是，不同鸟类在不同频率下似乎具有相似的动机（系列、颤音、咏叹调等）。例如，在图 4 中，两种不同的鸟类（布莱斯苇莺和亚洲杜鹃）共享一个动机模式。第三个问题是，像喇叭声和鸟类等外部声源在同一或不同频率上看起来可能相似（在动机方面）。如果我们能以某种方式将频率信息嵌入到梅尔频谱图中，我们希望能解决这一问题。接下来，我们讨论这种方法并将我们的讨论分为以下几部分：（1）声学事件检测，（2）特征工程梅尔频谱图，（3）通过三原色添加嵌入频率信息。最后，我们讨论模型架构。

2.1. 声学事件检测

我们使用具有 182 种分类的 BirdCLEF 2024 [1] 数据集。在其中，所有音频记录都被弱标记，即在录音级别而不是持续时间级别上存在标签。其持续时间可以在 3 秒到 30 分钟之间变化。首先，对音频进行去噪以分离鸟类声音活动的实例，并应用截止频率为 300Hz 的高通滤波器。我们发现低于此阈值的声学事件不构成任何显著活动。这导致录音中主要包含目标声学事件。去噪和高通滤波器降低了无关声音的能量水平。接下来，能量被计算为每帧中样本的平方绝对值的总和。高于平均能量的下降能量峰值时间点被计算出来（使用 `find_peaks` 包在 `scipy.signal library` 中）。然后在这些能量峰值周围包裹一个 5 秒窗口，完全包围构成声学事件的鸟类主题。然后进行多事件级别的例子分组，条件是接下来的低峰能量实例不应与前面的声学事件共享超过 50 % 时间（以防止声学事件重叠）。总体上，我们选择 5 个事件，每个持续时间为 5 秒（有时根据显著声学事件的存在可能事件数量更少）。在一个焦点录音中，在 5 个高于平均能量的声音事件中发现标记音频事件的概率被假定为 1，这意味着与 PAM 中设备记录整个声音而没有焦点相比，有意从特定兴趣区域捕捉声音。因此，我们的过程保证在 5 个挖掘的声音事件中包含主要标签声音。

[1]https://www.kaggle.com/competitions/birdclef-2024



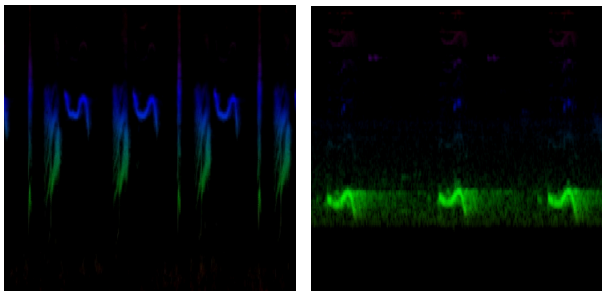
(a) *Blyth's Reed Warbler* (b) *Asian Koel*

Figure 4: 灰度梅尔频谱图中的两种鸟类共享一些主题模式

接下来，我们为检测到的声学事件创建一个梅尔频谱图。如本节开始时所述，梅尔频谱图是单通道的，代表录音中不同声学事件的音高变化或时间拉伸现象。像 Resnet [5] 、可变形卷积网络（DCNs）[6] 等计算机视觉模型在某种意义上是平移不变的，因此如果模式被共享，它们将为不同的声学事件预测相同的标签。或者，在共享模式的情况下，模型也会为次要标签分配较高的概率，从而使分类不太准确。因此，输出不能被完全依赖，我们需要在学习机制中以某种方式编码声音频率。为此，梅尔频谱图首先在 (0,1) 范围内归一化，然后在 log 尺度上缩放，然后再次在 (0,1) 范围内归一化。

如前所述，当将梅尔频谱图输入深度学习模型时，其失去了频率信息。为了解决这个问题，我们采取以下步骤。在图像中的任何梅尔频率区间的像素表示为 RGB 通道颜色比例 (1:1:1)。为了根据频率信息的变化区分梅尔频率区间的像素，梅尔频谱图在最低频率（f_min ）和最高频率（f_max ）之间被分为三个相等的区域，这些频率在创建梅尔频谱图时被初始化。因此，一个区域中的梅尔频率区间数量将为 $n\_bins = total\_bins/3$ ，其中 $total\_bins$ 是在梅尔频谱图创建时设置的参数。梅尔频率区间的第一个区域的频率范围始于 f_min Hz ，其主色像素将被映射到 `color_array` RG($1-t, t$) ，其中 t 定义为随着区间索引的增加，红色通道从纯红线性减少，而绿色通道线性增加，两种颜色同时以相同的量在每个区间中向上过渡。这种主色通道的添加为区域内所有的梅尔频率区间赋予了独特的次色。最后，梅尔频率区间的像素值与颜色数组 `color_array*pixel_value` 相乘。相同的操作在其他区域的梅尔频率区间上分别采用颜色数组 `color_array` GB($1-t, t$) 和 BR($1-t, t$) 进行。结果，我们得到了一种彩色化的梅尔频谱图，如图所示 5 。注意，彩色化可以被视为一种梅尔频谱图中频率信息编码的 近似 。因此，彩色化的梅尔频谱图可能有助于区分共享相同主题模式的两种不同鸟类，这将在实证部分中展示。

对于音频分类，我们使用 EfficientNetB0 架构，这是



(a) *Blyth's Reed Warbler* (b) *Asian Koel*

Figure 5: 两种鸟类声谱图显示共享一些母题模式。由于颜色化，深度学习模型现在可以区分这些母题。

一种 CNN，用于学习之后使用 AutoPool 层，如图所示。AutoPool 是一种为弱标记音频分类任务场景设计的池化机制，涉及多实例学习（MIL）。与传统的池化方法如最大池化或平均池化不同，AutoPool 引入了一种可训练的池化函数，该函数在训练过程中学习如何将实例级别的预测聚合为录音级别的预测。这种灵活性允许模型根据数据集特征自适应地在最大池化（突出主导信号）和平均池化（捕捉更广泛的模式）之间平衡。由于学习是基于弱标记数据进行的，而在测试期间，预测则在录音级别进行，因此 AutoPool 非常适合我们的任务。AutoPool 层将每个录音的 5 个实例的 logits 池化，然后通过 sigmoid 激活函数进行多类别多标签预测。对于实例少于五个的录音，用零图像通道填补剩余实例，在 AutoPool 之前对零图像通道应用二进制掩码。

多实例学习的自动池化概率聚合由以下公式给出。

其中，$\hat{P}(Y/x)$ 是类别 Y 给定实例 x 的概率，α 是在训练过程中与模型参数一起学习的标量参数，用于控制池化行为。特别地，$\alpha = 0$ 等于未加权均值，$\alpha = 1$ 等于软最大池化，$\alpha \to \infty$ 是一个最大操作符。注意，自动池化对于每个类别是分别进行的，以应对多标签问题。然而，我们没有为每个类别单独改变 α ，我们将这留给未来研究。

2.1.1. 损失函数

图 **??** 中的模型是针对二元交叉熵损失函数 (1) 进行优化的。

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\left[y_{ij}\log\hat{y}_{ij} + (1-y_{ij})\log\left(1-\hat{y}_{ij}\right)\right] \tag{1}$$

其中 N 是样本量，C 是类别数量。

我们利用 BirdCLEF 2024 数据进行拟议模型的训练和验证。它包含 182 种鸟类的 24459 段音频录音，其中我们只选取了 23920 段（可能有一些重复）。由于 BirdCLEF 竞赛的隐藏测试数据不可用，测试集是从给定的音频文件中准备的。我们过滤掉同时具有主要和次要标签的文件，因为次要标签是噪声且不可靠。我们丢弃了 1873 个文件。为了与现有文献中使用的测试数据保持一致，我们仅使用具有主要标签的录音。剩余的文件按 80:20 的比例划分用于训练和验证。

2.2. 基线

我们使用来自 BirdCLEF 2024 的获胜模型 [2] 作为基线，该模型本质上是一个 EfficientNET-B0，并对输入应用了一些增强方法，如水平切片、利用伪标签数据等。此外，我们移除了梅尔频谱图中的颜色化处理，以进行消融研究，观察颜色添加剂的效果。

[2]https://www.kaggle.com/competitions/birdclef-2024/discussion/512197

Table 1: 对比所提方法与 *BirdCLEF 2024* 获胜模型在无着色情况下的 *5* 折验证集上的性能。* 表示在 *Wilcoxon* 符号秩检验（单尾）中，与无着色模型相比有统计显著差异，$\alpha = 0.05$ 。

Model	Macro-F1	Macro ROC-AUC	CMAP
Winner model BirdCLEF	0.6371	0.9220	0.6915
Proposed Model w/o Colorization	0.6676	0.9765	0.7217
Proposed Model w/ Colorization	0.6833*	0.9797*	0.7374*

2.3. 训练和评估细节

正如在 §?? 中讨论的那样，模型采用 K 折交叉验证（在我们的情况下，K=5）策略进行训练。在验证期间，我们从验证集中选取前 5 个声学事件窗口进行输入，具体程序如 §2.1 所述。注意，相对于仅检测前 5 秒窗口以确定物种的存在（如 BirdCLEF 的获胜者所做，这可能包含或不包含 PAM 环境中实际物种），这种评估对于 PAM 录音（用于 BirdCLEF 隐藏集评估）更加现实。模型的优化初始学习率为 $3\ e^{-3}$ ，按余弦退火学习率调度器下降到 $1\ e^{-6}$ 。我们将批量大小设置为 90，训练轮数为 30。使用 AdamW 优化器来最小化损失函数。获胜模型是直接从获胜者提供的代码中借用的。我们使用相同的数据拆分来训练和测试获胜模型以进行公平比较。

2.4. 评估指标

宏观 ROC-AUC、宏观 F1 和类平均平均精度（CMAP）用于模型的性能评估。这些指标最适合处理具有不平衡类别的多类别、多标签问题。简而言之，CMAP 是以往 BirdCLEF 比赛中使用的每个类别精度分数的平均值 [7] 。然而，BirdCLEF 2024 使用了宏观 ROC-AUC。

2.5. 结果

如表 1 所示，我们可以观察到所提出的方法在所有指标上均优于获胜模型，分别在 F1 上提高 7.3 %，在 ROC-AUC 上提高 6.2 %，在 CMAP 上提高 6.6 %。有趣的是，我们在训练/推理过程中没有使用任何数据增强，而获胜模型使用了。为了观察所提出的频率嵌入的颜色添加剂的效果，我们进行了一项消融研究。如表 1 （第二行）所示，我们发现彩色处理在识别具有相似音型的鸟类物种方面是有效的。

3. 结论与未来工作

本研究通过多示例学习的视角研究鸟类分类问题。为了应对识别具有相似图案鸟种的问题，我们提出了频率嵌入的颜色添加剂的概念。实证结果显示，颜色化是在不同频段分类具有相似图案的鸟类的一种有效方法。可以继续的研究方向有多个。我们假设颜色化在与语音活动检测模块结合进行音频分割时，对于分类重叠的发声将更为有效。另一个有趣的研究可能是观察 α 参数在多类多标签问题中的效果。

4. 局限性

我们的研究仅限于识别主要标签。因此，我们的模型无法对音频记录中存在的未知鸟种进行分类。

5. References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, "Unsupervised sound separation using mixture invariant training," *Advances in neural information processing systems*, vol. 33, pp. 3846–3857, 2020.

[3] T. Denton, S. Wisdom, and J. R. Hershey, "Improving bird classification with unsupervised sound separation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 636–640.

[4] E. Birding, "Visualizing sound," 2024, accessed: 2024-12-28. [Online]. Available: http://earbirding.com/blog/specs/visualizing-sound

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[6] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 764–773.

[7] S. Kahl, F.-R. Stöter, H. Goëau, H. Glotin, R. Planque, W.-P. Vellinga, and A. Joly, "Overview of birdclef 2019: large-scale bird recognition in soundscapes," in *CLEF 2019-conference and labs of the evaluation forum*, vol. 2380, no. 256. CEUR, 2019.