arXiv:2302.08136v1 [cs.SD] 16 Feb 2023

基于注意力的分层多标签乐器分类方法

Zhi Zhong、Masato Hirano、Kazuki Shimada、Kazuya Tateishi、Shusuke Takahashi、Yuki Mitsufuji

索尼集团公司，日本东京

摘要

尽管音乐通常具有多标签属性，许多研究仍在简化的设置下探讨分层音乐标签，例如使用单标签数据。此外，目前尚缺乏一个能够描述多标签设置下各种联合训练方法的框架。为探讨上述问题，我们引入了分层多标签乐器分类任务。该任务提供了一种更贴近现实的设置，假设数据为多乐器真实音乐数据。我们总结并探讨了多种联合训练深度神经网络的分层方法，并将其置于深度学习与传统技术相结合的背景下进行考察。为在多标签设置下实现有效的联合训练，我们提出了两种方法来建模细粒度和粗粒度标签之间的联系：一种采用基于规则的分组最大池化，另一种则以数据驱动方式获得注意力机制。评估结果表明，与不进行联合训练的方法相比，所提方法具有优势。此外，通过可视化注意力图或参照固定规则，可以解释所提方法中的决策过程。

索引词——音乐标注，分层分类，多标签分类，乐器，注意力

1. 引言

多标签音乐标注是一项分类任务，旨在为给定的音乐作品预测多个语义标签。标签可以表示音乐的音乐流派、情绪和乐器。因此，这项任务对于音乐推荐或音乐检索等应用具有重要意义。将音乐标签组织成树状结构，即图 1 所示的层次结构，可以体现领域知识（哪些标签在音乐层面上相关），从而带来提升标注性能等益处 [1, 2, 3, 4]。尽管音乐标签数据集通常采用扁平层级结构 [5, 6, 7, 8],，音乐信息检索领域对分层标注的兴趣日益浓厚 [9]。

已有多项研究尝试解决分层音乐标注问题。Parmezan 等人研究了采用传统机器学习方法（不涉及深度学习）的分层音乐流派分类 [10]。也有少数研究在简化的问题设定下探讨了基于深度神经网络的层次化方法。例如，Garcia 等人针对单乐器数据研究了小样本乐器分类 [11], , 而 Nolasco 等人则利用单音数据研究了乐器表征学习 [12]。针对分层多标签音乐标注任务，Krause 等人报告了多种用于歌唱活动检测的基于深度神经网络的层次化方法 [13]。他们的讨论大多集中于分别训练多个深度神经网络，而非联合训练这些网络。

尽管真实音乐通常是复调且包含多种乐器，但许多研究都使用简化的数据探讨分层音乐标签。因此，基于深度神经网络的分层音乐标签在现实问题设定下仍未得到充分讨论。此外，目前缺乏一个框架来描述多标签设置下的各种联合训练方法。尽管使用多个独立模型是有效的，但已有研究表明，深度神经网络能够在联合训练过程中学习层次信息 [1, 2, 3, 4, 11, 12]。

本文采用单个深度神经网络的联合训练方法，研究层次化多标签音乐标注。为研究层次化多标签音乐标注，我们提出了多标签音乐乐器分类任务，其中的乐器按照两级层次结构组织。本文的贡献如下。首先，该任务提供了一个更贴近现实的场景，我们研究的是真实音乐，而真实音乐通常具有复调、多乐器和音乐类型多样等特点。其次，我们在一个类似于传统分层方法分类框架的框架下，对深度神经网络的各种联合训练方法进行分类和探索，以促进对深度学习与传统技术融合的进一步研究。最后，为了在多标签设置下实现有效的联合训练，我们提出了 ResAtt 和分组最大池化（ GMP ），通过应用残差注意力层或最大池化操作来建模细粒度和粗粒度标签之间的关联。

2. 相关工作

在音频领域，现有的分层分类研究主要集中于声音事件检测任务。在 [1], 中，粗粒度标签与细粒度标签之间的关联被表述为分组求和池化。该表述要求细粒度预测经过 softmax 激活归一化，因此侧重于单标签任务。Zharmagambeto 等人将深度神经网络与决策树结合起来 [4], , 以鼓励深度神经网络学习包含层次信息的表征。然而，推理阶段使用了独立分类器，因此该分类器未能通过层次信息进行优化。

Nolasco 等人研究了 Nsynth数据集 [14] 中乐器的层次化度量学习 [12] （单个乐器演奏单个音符）。文献 [11]则在小样本学习设置下研究了复调版本的乐器任务，所用的单乐器数据取自 MedleyDB数据集的分轨音轨[15]。剔除“Drum”等不够细粒度的标签后，根据音乐领域采用的乐器分类方式构建层次结构 [16]。这一思路启发我们将自己提出的基于音色的层次结构（见图 1）引入实验所用的数据集中。

本文从歌唱活动检测的角度讨论层次化多标签音乐标注。为建立单个深度神经网络所生成的粗粒度预测与细粒度预测之间的联系，本文引入了损失项，以提升检测性能。此外，本文还介绍了一个用于描述各种基于深度神经网络的层次化方法的框架；然而，许多研究都致力于采用独立的多模型方法，而非联合训练这些模型。

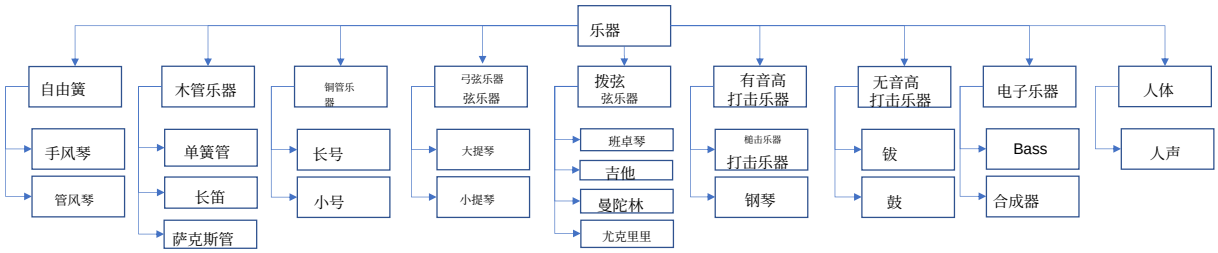
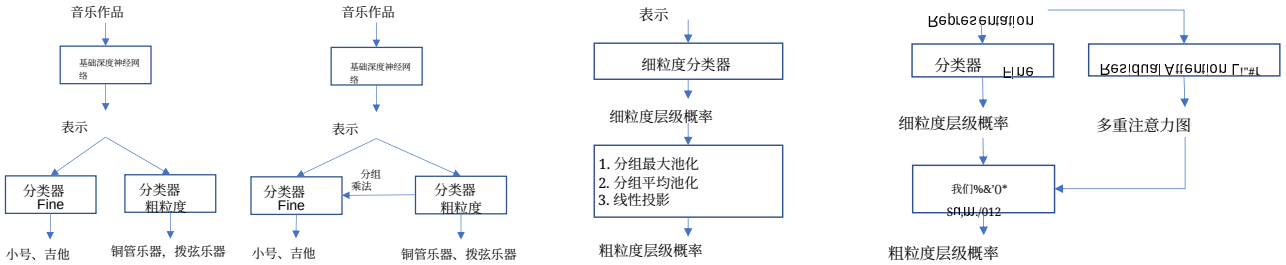


图 1. OpenMIC 数据集中的诱导式基调两级乐器层级



（a）逐层方法　(b) 自顶向下方法　(c) 自底向上方法　（d）ResAtt：提出的自底向上方法

图2。DNN 的联合训练方法

3. 提出的方法

假设基础深度神经网络提取了音乐作品的 N_{dim} 维表示 $\mathbf{Z} \in \mathbb{R}^{N_{\text{dim}}}$。随后，分类器通过 Sigmoid 激活将该表示映射为细粒度层级概率 $\mathbf{P}_{\text{fine}} \in \mathbb{R}^{N_{\text{fine}} \times 1}$。使用逐标签阈值将 $\mathbf{P}_{\text{fine}}$ 转换为二进制值，以执行二元分类。这里假设采用两层（细粒度／粗粒度）层次结构。

3.1. 联合训练的层次化方法

本文的主要 focus 是对用于层次化多标签分类的深度神经网络进行联合训练。我们首先按照 [13, 17], 中所述传统层次化技术的分类方式，总结各种相关方法，并展示深度学习如何与传统技术相结合。

逐层方法。传统上，会针对层次结构中的每个级别分别准备一个模型 [17, 13]。 [13], 将这一思想应用于联合训练框架，其结构如图2（a）所示。尽管该方法有助于基础深度神经网络学习层次信息，但分类器在训练过程中仍彼此独立。在 [13], 中，引入了两个损失项，以利用层次信息优化这些分类器。

自顶向下方法。传统上，首先训练一个模型对粗粒度标签进行分类，然后针对每个粗粒度标签分别准备模型，以分类其子标签（细粒度标签） [17, 10, 13]。我们将其概括为“先粗后细”。为了将该方法应用于深度神经网络的联合训练，我们在基础深度神经网络之后连接了一个简化的软决策树（SDT） [18]，其类似于 [4]。在软决策树中，细粒度概率是叶节点概率与其父节点（粗粒度概率）的乘积 [18],，在图2（b）中称为粗粒度标签与其子细粒度标签之间的分组乘法。由于假设任务为多标签任务，软决策树中的Softmax激活被Sigmoid取代。不同于 [4],，我们使用软决策树进行推理，因为软决策树在训练过程中会被联合优化。然而，在自顶向下方法中，粗粒度级别上的错误很难纠正，这可能会影响其细粒度层级性能 [13]。

自底向上方法。在传统的自底向上方法中，只使用细粒度标签训练非层次化（扁平）模型。推理时，通过预定义规则 [13, 17] 汇总细粒度预测，得出粗粒度标签。我们将其概括为“先细粒度，后粗粒度”。传统自底向上方法的核心问题在于，模型并未与“自底向上规则”联合优化，因此性能并非最优。另一个主要缺点与自顶向下方法类似：细粒度错误会传播到粗粒度级别 [13, 17]。

自底向上聚合的一条常见规则是：当音乐样本被赋予细粒度标签时，也会为该音乐赋予其父级粗粒度标签。例如，如果某个音乐样本被标注为“长笛”，那么它也会被标注为“木管乐器”。我们将这条规则称为分组最大池化（GMP），因为该规则等同于对不同组的子级细粒度标签应用最大池化。显然，层次化数据集中的细粒度和粗粒度标签也满足相同的规则 [11, 13, 19]。

我们强调结合联合训练的自底向上方法这一概念，即通过聚合规则对深度神经网络进行联合优化。鉴于数据集中的细粒度和粗粒度标签满足类似于最大池化的规则，我们提出在训练期间应用 GMP（图2（c）），以使深度神经网络了解数据集中的层次结构，并提升标注性能。

3.2. ResAtt：基于注意力的自底向上方法

我们还提出了 ResAtt，它通过注意力机制对细粒度和粗粒度标签之间的关联进行建模。ResAtt 可以利用元素值介于 0% 至 100% 之间的注意力图，告知系统对于特定粗粒度标签，哪些细粒度标签重要、哪些不重要。这一想法源于这样一种认识：最大池化等同于一种操作，它会生成 0/1 二值注意力图，从输入向量中选出最重要的元素。因此，这两种提出的方法都可以视为基于注意力的方法。

如图2 (d)所示，使用残差注意力层为每个粗粒度标签生成注意力图，记为 $\mathbf{W} \in \mathbb{R}^{N_{\text{fine}} \times N_{\text{粗粒度}}}$；将注意力图应用于细粒度预测，最终通过以下公式预测粗粒度标签：

$$\mathbf{P}_{\text{coarse}} = \mathbf{W}^T \cdot \mathbf{P}_{\text{fine}} \in \mathbb{R}^{N_{\text{coarse}} \times 1}, \tag{1}$$

其中，T是矩阵的转置。对注意力图 W 的 N_{fine} 维应用 softmax 激活，以确保得到的粗粒度概率保持合理（低于 100%）。

在 ResAtt 中，要获得粗粒度标签的预测，首先需要细粒度分类器给出高质量的结果。这有助于在训练过程中联合优化分类器和基础深度神经网络。与通过树结构或最大池化规则获知数据集层次结构的自顶向下方法或采用 GMP 的联合训练方法不同，ResAtt 无法获取这类先验知识，而是以数据驱动的方式探索合适的聚合规则。在第5节中，我们将讨论这一特性如何帮助防止细粒度错误向粗粒度级别传播。

按照 [1, 11], , 整体二元交叉熵损失表示为分层损失的加权和，即

$$L_{\text{BCE}} = \lambda L_{\text{BCE}}^{\text{fine}} + (1-\lambda) L_{\text{BCE}}^{\text{coarse}}, \tag{2}$$

其中，λ 是细粒度标签的权重。

4. 评估

4.1. 基于音调的层级数据集

OpenMIC [20] 是一个乐器分类数据集，其任务与实际应用场景十分接近：从 FMA数据集 [19] 中选取不同音乐流派的10秒音乐片段，并以多标签方式对20种乐器进行标注。虽然官方标注仅包含细粒度标签，但我们根据乐器的音调属性构建了一个两层乐器层次结构（图1）。这一预处理与 [11]中描述的方法类似。不过，我们的层次结构与 [11] 不同：我们将被认为不够“细粒度”的标注（如“键盘打击乐器”或“鼓”）也纳入层次结构 [11]，因此可以使用 OpenMIC 提供的所有音乐片段。

由于没有用于超参数调优的官方验证集，我们按照 [21] 中的做法，从训练集中抽取 15% 的数据用于验证。使用 scikit-multilearn 库 [22] 进行分层抽样。

OpenMIC 数据集中有许多乐器尚未标注，因此发布时附带了一个掩码文件，告知用户某个音轨中具体检查了哪些乐器。我们遵循常见做法，在训练期间使用该掩码文件计算损失，并在评估期间计算评估指标 [21, 23]。损失计算的伪代码可写为 Loss $= L_{\text{BCE}}$(predictions[mask],label[mask])。

4.2. 实验

我们使用在 AudioSet [24] 上预训练的 CNN14 架构作为基础深度神经网络，其中 $N_{\text{dim}} = 2048$，细粒度分类器是一个单独的线性层。原始音乐数据被转换为 16kHz 采样率的单声道音频，随后通过短时傅里叶变换（使用 32-ms 汉宁窗和 10-ms 帧移）转换为 64 个频带的梅尔频谱图（频率范围： [50Hz，8kHz]）。输入音频长度与 OpenMIC 中的片段长度相同。

我们评估了多种方法，并与所提出的方法进行比较；这些方法的模型大小基本保持在同一水平。我们使用八个不同的随机种子训练以下方法，并计算其评估指标。扁平基线。我们训练一个 CNN14 模型，作为细粒度任务的扁平基线。在推理阶段应用 GMP，生成该模型的粗粒度预测（ [4, 13]中采用的传统自底向上方法）。我们将 [23] 中报告的细粒度结果与我们的方法进行比较基线。逐层方法。我们评估图2（a）所示的 DNN，其中包含两个并行线性层。我们将 [13] 中提出的损失项引入架构，并使用式（2）中的二元交叉熵损失训练 DNN。自顶向下方法。我们评估图2（b）所示、带有软决策树分类器的 DNN，其中树的每一层都是一个单独的线性层。联合训练的自底向上方法。我们评估图2（d）中的 ResAtt，其注意力层通过重塑线性层的输出实现，计算开销与逐层方法或自顶向下方法一样低。我们也评估了所提出的采用 GMP 的联合训练方法。为了与 ResAtt 比较，我们用线性投影（LP）层替换注意力机制。同样，我们用平均池化替换最大池化，构成分组平均池化（GAP）方法，以便与采用 GMP 的联合训练方法进行比较。我们按照图2（c）联合训练 GAP 和 LP。

训练过程中，批量大小为 16，最大学习率为 1e-4。前 5 个训练轮次采用线性递增的学习率（线性预热），以提高训练稳定性。使用带有 1e-4 权重衰减的 Adam优化器 [25]。我们使用 SpecAugment [26] 进行数据增强。针对损失函数（公式）中的权重 λ 2 $\lambda \in$ 进行网格搜索，其中。选择细粒度标签验证损失最低的设置进行评估。 $[0.70, 0.75, 0.80, 0.85, 0.90]$

4.3. 评估指标

为进行客观评估，我们采用 ROC-AUC、PR-AUC（也称为平均精度均值）和 F1分数的宏平均值。所有评估指标均使用 scikit-learn库 [27]计算。F1分数基于 0/1 二元预测计算，这些预测由在验证集上优化的阈值产生。

5. 结果

表 1 展示了不同随机种子下评估结果的平均值，其中的数字表示不同的设计选择。

我们的细粒度分类模型在 [23] 中优于 PaSST-S。这表明，我们对基础深度神经网络的选择以及训练设置是可行的。

结合 [13] 中损失项的逐层方法大幅提升了粗粒度层级性能，并且尤其在 F1分数 方面改善了细粒度层级性能。然而，该模型内部的决策过程难以解释。与此同时，由于分类器 [18] 具有树结构，自顶向下方法具备可解释性。在粗粒度级别上，自顶向下方法的性能与逐层方法相近；但在细粒度级别上，该方法导致 PR-AUC 值下降。如第 3.1 节所述，其细粒度层级性能可能受到了粗粒度层级性能的限制。

所提出的采用 GMP 的联合训练方法与我们的基线之间唯一的区别在于，所提出的方法通过联合训练，依据 GMP 规则优化深度神经网络。采用 GMP 的联合训练方法可通过 GMP 规则进行解释，例如，“木管乐器”之所以被预测，是因为“长笛”“单簧管”或“萨克斯管”中至少有一种被以较高概率预测。如第 3.1 节所述，音色基础层次数据集中的细粒度和粗粒度标注满足类似 GMP 的规则，这解释了 GMP 联合训练带来的显著性能提升。将最大池化替换为平均池化操作后，便得到 GAP 方法。由于在该数据集中，粗粒度标签并非其子级细粒度标签的平均值，因此使用 GAP 对深度神经网络进行联合训练会导致性能下降。

表 1. 基于音色的层次结构上的结果（%）

		细粒度级别			粗粒度级别		
注释	方法	ROC-AUC	PR-AUC	F1分数	ROC-AUC	PR-AUC	F1分数
	PaSST-S [23]	-	84.3	-	n/a	n/a	n/a
基线	扁平细粒度分类与自底向上的推理阶段	91.7	85.6	83.3	92.4	89.6	84.4
图2 (a)	逐层方法，在 [13] 中包含损失项（我们的调优）	91.9	85.8	**83.7**	93.0	90.4	85.2
图2 (b)	自顶向下方法	91.9	85.4	83.6	93.0	90.4	85.3
	采用联合训练的自底向上方法						
图2 (c)	1. 分组平均池化（GAP）	91.7	85.6	83.2	91.7	87.6	84.6
图2 (c)	2. 线性投影（LP）	91.8	85.8	83.6	75.3	70.9	67.6
图2 (c)	3. 分组最大池化（GMP）（所提出的方法）	91.9	85.9	83.6	**93.1**	**90.7**	85.4
图2（d）	4. ResAtt（所提出的方法）	**92.0**	**86.0**	**83.7**	**93.1**	**90.7**	**85.5**

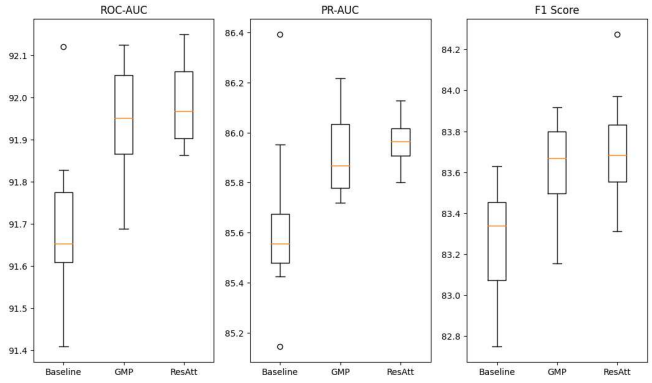


图3. 基线模型和所提出模型的细粒度指标箱线图。橙色线表示中位数

与采用 GMP 的联合训练方法不同，ResAtt 在训练期间不考虑数据集的层次结构，而是尝试仅从标签中提取聚合规则。ResAtt 的得分与采用 GMP 的联合训练方法相当，或略高一些。LP 方法使用可学习的投影层，而非注意力层，但未能预测粗粒度标签。这是因为线性层无法针对不同输入灵活调整其投影规则，说明在采用联合训练的自底向上方法中使用注意力机制至关重要。

由于细粒度指标相较于基线的提升绝对值较小，我们通过箱线图（图3）比较所提出的方法与基线。较高的中位数和箱体表明了它们相对于基线的优势。尽管ResAtt的中位数略高，但采用GMP的联合训练方法与ResAtt之间的差异并不显著。

图4展示了ResAtt生成的注意力图样例及其逐标签概率。所有注意力图都与GMP操作相似，但会随输入音乐而变化，这表明ResAtt无需先验知识便提取出了聚合规则。在下方的例子中，尽管细粒度级别的“钢琴”被错误地预测为83%，ResAtt却通过降低对“钢琴”的注意力，没有将这一错误传递到粗粒度级别。这表明ResAtt能够学习难以用固定规则描述的灵活聚合方式，在某些情况下甚至能够阻止错误传播到粗粒度级别。采用GMP的联合训练方法无法实现此类行为，因为该方法会累积细粒度级别中的错误[13, 17], 这可能是ResAtt在粗粒度级别的F1分数略高的原因。通过注意力图可视化，可以解释ResAtt内部的决策过程。

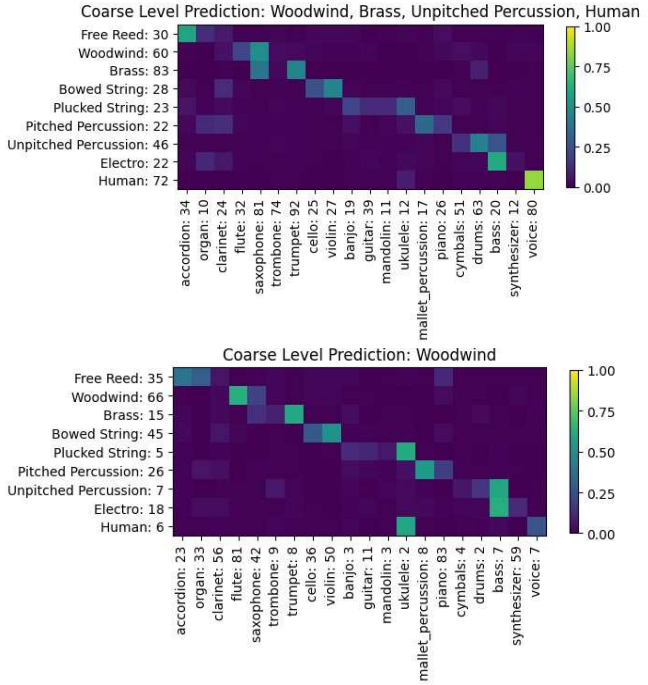


图 4. 注意力图。顶部列出粗粒度标签预测；各标签名称旁列出逐标签概率（%）。

6. 结论与未来工作

我们以分层音乐标注为案例研究，探讨了分层多标签乐器分类。我们将分层乐器分类扩展到多标签设置和真实音乐数据，并构建了诱导的基于音色的层次结构。在深度学习与传统技术融合的背景下，我们总结了多种联合训练深度神经网络的分层方法。为了在多标签设置中实现有效的联合训练，我们提出了两种建模细粒度标签与粗粒度标签之间联系的方法：一种采用以数据驱动方式获得的注意力机制，另一种采用基于规则的分组最大池化，并将其解释为二元注意力。评估结果表明，所提出的方法，尤其是 ResAtt，作为采用联合训练的自底向上方法，具有良好前景。除性能表现外，ResAtt 还能学习难以用固定规则描述的灵活聚合方式，甚至可以防止错误传播。通过可视化注意力图，可以增强 ResAtt 的可解释性。未来工作包括将现有方法扩展到音乐标注中的其他任务。

7. 参考文献

[1] Abelino Jimenez、Benjamin Elizalde 和 Bhiksha Raj，“使用基于本体的神经网络进行声音事件分类”，载于 NeurIPS 2018 论文集，2018 年，第 9 卷。[2] Arindam J ati、Naveen Kumar、Ruxin Chen 和 Panayiotis Georgiou，“树结构标签空间上的层级感知损失函数用于音频事件检测”，载于 ICASSP 2019 论文集。IEEE，2019 年，第 6–10 页。[3] Yiwei Sun 和 Shabnam Ghaffarzadegan，“用于音频事件分类的本体感知框架”，载于 ICASSP 2020 论文集。IEEE，2020 年，第 321–325 页。[4] Arman Zharmagambetov、Qingming Tang、Chieh-Chi Kao、Qin Zhang、Ming Sun、Viktor Rozgic、Jasha Droppo 和 Chao Wang，“使用树结构本体改进声学事件分类的表征学习”，载于 ICASSP 2022 论文集。IEEE，2022 年，第 321–325 页。[5] Edith Law、Kris West、Michael I Mandel、Mert Bay 和 J Stephen Downie，“使用游戏评估算法：音乐标注案例”，载于 ISMIR2009 论文集，2009 年，第 387–392 页。[6] Thierry Bertin-Mahieux、Daniel P.W. Ellis、Brian Whitman 和 Paul Lamere，“百万歌曲数据集”，载于 ISMIR 2011 论文集，2011 年。[7] Dmitry Bogdanov、Minz Won、Philip Tovstogan、Alastair Porter 和 Xavier Serra，“用于自动音乐标注的 MTG-Jamendo 数据集”，载于机器学习音乐发现研讨会，ICML 2019，长滩，加利福尼亚州，美国，2019 年。[8] Minz Won、Andres Ferraro、D mitry Bogdanov 和 Xavier Serra，“基于 CNN 的自动音乐标注模型评估”，载于第17届声音与音乐计算会议论文集，2020 年。[9] Minz Won、Janne Spijkervet 和 Keunwoo Choi，《音乐分类：超越监督学习，迈向现实世界应用》，https://music-classification.github.io/tutorial，2021年11月。[10] Antonio Rafael Sabino Parmezan、Diego Furtado Silva 和 Gustavo EAPA Batista，“分层音乐流派分类的局部方法组合”，载于 ISMIR 2020 论文集，2020 年，第 740–747 页。[11] Hugo Flores Garcia、Aldo Aguilar、Ethan Manilow 和 Bryan Pardo，“利用层次结构进行少样本乐器识别”，载于 ISMIR 2021 论文集，2021 年。[12] Inˆes Nolasco 和 Dan Stowell，“用于学习层次表征的基于排序的损失”，载于 ICASSP 2022 论文集。IEEE，2022 年，第 3623–3627 页。[13] Michael Krause 和 Meinard M¨uller，“复杂音乐录音中的歌唱活动、性别和类型的分层分类”，载于 ICASSP 2022 论文集。IEEE，2022 年，第 406–410 页。[14] Jesse Engel、Cinjon Resnick、Adam Roberts、Sander Dieleman、Douglas Eck、Karen Simonyan 和 Mohammad Norouzi，“使用 WaveNet 自动编码器进行音乐音符的神经音频合成”，2017 年。[15] Rachel M Bittner、Justin Salamon、Mike Tierney、Matthias Mauch、Chris Cannam 和 Juan Pablo Bello，“MedleyDB：用于注释密集型音乐信息检索研究的多轨数据集”，载于 ISMIR 2014 论文集，2014 年，第 14 卷，第 155–160 页。

[16] Erich M. von Hornbostel 和 Curt Sachs，“乐器分类：由 Anthony Baines 和 Klaus P. Wachsmann 从德文原文翻译”，《加尔平学会期刊》，第 14 卷，第 3–29 页，1961 年。[17] Carlos N Silla 和 Alex A Freitas，“不同应用领域中的分层分类综述”，《数据挖掘与知识发现》，第 22 卷，第 1 期，第 31–72 页，2011 年。[18] Alvin Wan、Lisa Dunlap、Daniel Ho、Jihan Yin、Scott Lee、Henry Jin、Suzanne Petryk、Sarah Adel Bargal 和 Joseph E. Gonzalez，“Nbdt：神经网络支持的决策树”，载于 ICLR 2021，2021 年。[19] Micha¨el Defferrard、Kirell Benzi、Pierre Vandergheynst 和 Xavier Bresson，“FMA：用于音乐分析的数据集”，载于 ISMIR 2017 会议论文集，2017 年。[20] Eric Humphrey、Simon Durand 和 Brian McFee，“Openmic-2018：用于多乐器识别的开放数据集”，载于 ISMIR 2018 会议论文集，2018 年，第 438–444 页。[21] Siddharth Gururani、Mohit Sharma 和 Alexander Lerch，“用于乐器识别的注意力机制”，载于 ISMIR 2019 会议论文集，2019 年。[22] P. Szyma´nski 和 T. Kajdanowicz，“用于执行多标签分类的基于 scikit 的 Python 环境 ”，ArXiv 电子预印本，2017 年 2 月。[23] Khaled Koutini、Jan Schl¨uter、Hamid Eghbal-zadeh 和 Gerhard Widmer，“通过 patchout 高效训练音频变换器”，载于 Interspeech 2022 会议论文集，2022 年。[24] Qiuqiang Kong 、Yin Cao、Turab Iqbal、Yuxuan Wang、Wenwu Wang 和 Mark D Plumbley，“PANNs：用于音频模式识别的大规模预训练音频神经网络”，IEEE/ACM TASLP，第 28 卷，第 2880–2894 页，2020 年。[25] Diederik P Kingma 和 Jimmy Ba，“Adam：一种随机优化方法”，arXiv 预印本 arXiv:1412 6980，2014 年。。[26] Daniel S Park、William Chan、Yu Zhang、Chung-Cheng Chiu、Barret Zoph、Ekin D Cubuk 和 Quoc V Le，“Specaugment：一种简单的数据增强方法，用于自动语音识别”，载于 Interspeech 2019，奥地利格拉茨，2019 年。[27] F. Pedregosa、G. Varoquaux、A. Gramfort、V. Michel、B. Thirion、O. Grisel、M. Blondel、P. Prettenhofer、R. Weiss、V. Dubourg、J. Vanderplas、A. Passos、D. Cournapeau、M. Brucher、M. Perrot 和 E. Duchesnay，“Scikit-learn：用于机器学习的 Python”，《机器学习研究期刊》，第 12 卷，第 2825–2830 页，2011 年。