基于卷积神经网络的音频场景分类方法研究

**摘要**

1. **绪论**
   1. 课题背景及研究的目的及意义

声音，一种传递信息的媒介。我们能从声音中提取到不少有用的信息。伴随着人工智能技术的发展，借助机器辅助人们从声音中提取出信息成为了人工智能领域一大研究方向。具体来说，音频处理任务中包括自动音乐标记、音乐检索推荐、语音识别、音频指纹、音频场景识别等应用场景。本文主要研究音频场景识别。

音频场景识别的应用场景包括设计上下文感知服务（Adams, Want，1994），智能可穿戴设备（Xu, Li, Lee 2008），机器人导航系统（Chu, Narayanan, Kuo, Matari，2006）和音频归档管理（Landone, Harrop, Reiss，2008）。此外，智能个人助理也是一个受到音频场景识别推动的领域。 智能个人助理是通过分析各种输入数据—包括音频，图像，用户输入或位置，天气和个人时间表等上下文信息，自动进行推荐和执行操作的软件代理。当今智能个人助理服务的代表有Google的Google Now、微软的Cortana、Apple的Siri以及亚马逊的Alexa。这些服务从环境音频中提取上下文信息，可以向用户自动推荐具有价值的信息，是一种极具实用价值的人工智能应用。

音频场景识别的主要目标是通过分析声音使设备能够理解其环境。实现过程为：先对采集到的音频信号进行预处理，再从中提取用于区分环境的有用特征，最后根据这些特征进行分类。

此前音频场景分类基于将通用分类器（高斯混合模型，支持向量机，隐马尔可夫模型）应用于手动提取的特征，例如梅尔频率倒谱系数。近年来，得益于计算机速度的提升与深度学习的快速发展，人们逐渐意识到，可以尝试用深度学习的自动特征提取的特性来代替以往低效的手工提取。正如“深度学习”一词所表明的那样，该方法通过使用非线性模块堆叠多个层来进行低层数据的高级表示。有几种深度学习体系结构的变体，卷积神经网络其中的一种，由于其在学习独特的局部特征方面的优越性能，被广泛用于图像分类、语音识别、自然语言处理。与其他深度学习结构相比，卷积神经网络在图像和[语音识别](https://zh.wikipedia.org/wiki/%E8%AF%AD%E9%9F%B3%E8%AF%86%E5%88%AB)方面能够给出更好的结果。相比较其他深度、前馈神经网络，卷积神经网络即使在有限的数据集和简单的数据增强下也可以有效应用于环境声音分类任务。更重要的是，可用数据集规模的显著增加很可能大大提高训练模型的性能。得益于卷积神经网络对数据集的利用程度高及高效的类别学习特性，可以看出卷积神经网络对音频场景分类任务有很高的价值。

* 1. 国内外研究现状

早在1997年，MIT媒体实验室就已经展开了音频场景识别的工作。在研究起步时期，识别率不甚理想。而随着如今智能设备大量涌现，优秀的计算机能力与深度学习技术的发展共同推动了该领域的研究进程。基于本文的需要，下面分别从音频特征提取、卷积神经网络、音频场景识别三个方面介绍国内外研究现状。

* + 1. 音频特征提取研究现状

音频场景识别系统中采用了几类音频特征。在这里，我们列出它们，在音频场景识别的背景下阐述它们的基本原理。

1）低级基于时间和基于频率的音频描述符：几个ASC系统，采用可以从时域中的信号或其傅立叶变换容易地计算的特征。这些包括（以及其中）过零率，其测量信号内的符号变化的平均速率，并且与单声道声音的主频率相关;光谱质心，测量光谱的质心，它与亮度的感知有关;以及频谱滚降，它识别频率高于设定阈值的频率。

2）频带能量特征（能量/频率）：各种ASC系统使用的这类特征是通过在指定频带上对幅度谱或功率谱进行积分来计算的。得到的系数测量不同子带内存在的能量的量，并且还可以表示为子带能量与总能量之间的比率，以编码信号中最突出的频率区域。

3）听觉滤波器组：能量/频率特征的进一步发展在于通过滤波器组分析音频帧，其模仿人类听觉系统的响应。为此，Sawhney和Maes使用Gammatone过滤器，Clarkson等。相反，计算Mel级滤波器组系数（MFC），而Patil和Elahili使用所谓的听觉谱图。

4）倒谱特征：MFCC是倒谱特征的一个例子，也许是ASC中最常用的特征。它们是通过计算MFC对数的离散余弦变换（DCT）获得的。名称cepstral是光谱的字谜，表示通过对信号频谱应用傅立叶相关变换来计算这类特征。倒谱特征捕获声音的频谱包络，从而总结其粗糙的光谱内容。

5）空间特征：如果使用多个麦克风记录声景，则可以从不同声道提取特征以捕获声学场景的属性。

在立体声录音的情况下，流行的特征包括测量在录制声源时在左右声道之间发生的相对延迟的耳间时间差（ITD）;和耳间水平差（ILD）测量通道之间的幅度变化。

ITD和ILD都与立体声场中声源的位置相关联。 Nogueira等。包括ASC系统中的空间特征。

6）发声特征：每当认为信号包含谐波分量时，可以估计基频f或一组基频，并且可以定义特征组以测量这些估计的特性。在ASC的情况下，谐波分量可能对应于音频场景内发生的特定事件，并且它们的识别可以帮助区分不同的场景。盖格等人。采用与系统中每帧的基频相关的发声功能。 Krijnders和Holt提出的方法基于提取音调拟合特征，即从音频信号的感知动机表示导出的一系列发声特征。首先，计算所谓的耳蜗图以提供受人耳蜗特性启发的声学场景的时频表示。然后，评估每个时频区域的音调以识别声学场景中的音调事件，从而产生音调拟合特征向量。

7）线性预测系数（LPC）：这类特征已被用于分析被模拟为自回归过程的语音信号。在自回归模型中，给定时刻t的信号s的样本表示为L个先前时刻的样本的线性组合：

(1)

其中组合系数确定了模型参数和是一个剩余项。 LPCS的值与建模信号的频谱包络之间存在映射，因此编码关于声音的一般频谱特性的信息。 Eronen等人在他们提出的方法中采用了LPC特征。

8）参数逼近特征：自回归模型是近似模型的一种特殊情况，其中信号s表示为来自集合的J基函数的线性组合。

. (2)

9）无监督学习特征：模型（2）假设先验地定义了一组基函数来分析信号。或者，可以从数据或从已经以无人监督的方式提取的其他特征中学习基础。 Nam等人。采用稀疏受限玻尔兹曼机器（SRBM）自适应地学习训练数据MFCC的特征.SRBM是一种神经网络，已被证明可以从输入图像中学习基函数，类似于视觉受体构建的表示的属性。人脑。在ASC的上下文中，SRBM自适应地编码训练信号的频谱的基本属性，并返回从MFCC学习的一系列特征，以及用于确定包含重要声学事件的时间段的激活函数。

10）矩阵因子分解方法：用于音频应用的矩阵因子分解的目标是将声学信号的频谱图描述为捕获典型或显着频谱元素的基本函数的线性组合，因此是一类无监督学习特征。

证明使用矩阵因子分类进行分类的主要直觉是，在声学场景识别中重要的事件的签名应该在基本函数中编码，从而导致判别性学习。

Cauchi采用非负矩阵因子分解（NMF）和Benetos等人。在他们提出的算法中使用概率潜在分量分析。请注意，矩阵因子分解还输出一组激活函数，这些激活函数及时编码基本函数的贡献，从而对整个音景的属性进行建模。因此，可以考虑这类技术来联合估计局部和全局参数。

11）图像处理特征：Rakotomamonjy和Gasso为ASC设计了一种算法，其特征提取功能包括以下操作。首先，使用恒定Q变换处理对应于每个训练场景的音频信号，该恒定Q变换返回具有对数间隔频带的频率表示。

然后，通过内插相邻时频区，从恒定Q表示获得512×512像素灰度图像。最后，通过计算局部梯度直方图的矩阵从图像中提取特征。这是通过将图像划分为局部斑块，通过定义一组空间取向方向，以及通过计算呈现每个取向的边缘的出现来获得的。注意，在这种情况下，特征的矢量不是从帧中独立提取的，而是从常数Q变换的时频片中提取的。

12）事件检测和声学单元描述符：Heit-tola等人提出了一种ASC系统，该系统根据信号中检测到的事件直方图对声景进行分类。在训练阶段，使用手动注释事件（例如“汽车喇叭”，“掌声”或“篮球”）的出现来导出每个场景类别的模型。在测试阶段，HMMS用于识别未标记记录内的事件，并定义与从训练数据导出的直方图进行比较的直方图。该系统代表了包括特征，统计学习和决策标准的通用框架的替代方案，因为它基本上同时执行事件检测和ASC。但是，出于本教程的目的，声学事件可以被认为是高级特征，其统计特性由直方图描述。

Chauduri等人采用类似的策略来学习声学单元描述符（AUDS）并对YouTube多媒体数据进行分类。 AUDS使用HMMS建模，用于将录音转录为一系列事件。假设转录由N-克语言模型生成，其参数在不同的音景类别上进行训练。因此，在测试阶段期间未标记记录的转录按照最大似然标准进行分类。

* + 1. 卷积神经网络研究现状
    2. 音频场景识别研究现状
  1. 研究内容

1. **基于 MFCC 和 KNN 的场景识别基线系统**
2. **基于MFCC 和CNN的场景识别系统**
3. **实验结果分析及对比**

**结论**

**参考文献**

**致谢**