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
    2. 卷积神经网络研究现状
    3. 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**