基于卷积神经网络的音频场景分类方法研究

# 第1章 引言

## 1.1 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。

音频场景识别属于计算听觉场景分析(CASA)的子领域，其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。

## 1.2 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，优秀的计算机性能与深度学习技术的发展共同推动了该领域的研究进程。从过程来看，音频场景识别通常需要进行预处理、特征提取、分类器分类这一系列流程。因此，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。

### 1.2.1 音频特征提取研究现状

许多关于音频场景识别的工作都致力于特征提取部分，原因在于良好的特征简化了分类器的设计，而识别能力欠佳的特征很难用任何分类器来补偿。此前受到语音识别领域的推动，有许多特征已经被投入到广泛的研究与应用之中。下面将介绍几种常用特征，并简述其原理：

（1）简单时频特征：(Eronen,2006)及（Malkin,Waibel,2005）等人设计的音频场景分类系统中，采用了这类特征，其特点为可以通过简单时域计算或通过傅里叶变换得到。其中包括过零率：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）：(Eronen,2006)等人中使用的这类特征，是通过在指定频带上对幅度谱或功率谱进行积分得到的。得到的系数用来衡量不同子带内存在的能量，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/频率特征的进一步应用在于通过滤波器组分析音频帧，以模仿人类听觉系统的响应。它通过了一组带通滤波器，输出具有一定中心频率的子带信号。其中，（Sawhney和Maes，1997）使用Gammatone过滤器，（Clarkson等，1998）。采用了Mel级滤波器组系数（MFC），而（Patil和Elahili，2002）使用了听觉谱图。

（4）倒谱特征：倒谱特征是为了某些时候便于计算，将原信号的频谱转化为类似分贝的单位，再对其做逆傅里叶变换，将其视为一种新信号处理。MFCC则为音频场景识别中最常用的倒谱特征之一。其得以大量应用的主要原因就在于MFCC的频带划分是在梅尔刻度上等距划分的，它相比于正常对数倒谱中的线性间隔频带更接近人类的听觉系统。在之后的章节中将详细介绍该特征的原理以及提取过程。

（5）发声特征：每当认为信号中包含谐波分量时，可以估计其基频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。在音频场景识别的情况下，谐波分量可能与音频场景内发生的特定事件相对应，且可以通过其识别确定不同的音频场景。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，2013）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。

（6）线性预测系数（LPC）：线性预测是根据[线性预测](https://zh.wikipedia.org/wiki/%E7%BA%BF%E6%80%A7%E9%A2%84%E6%B5%8B)模型的信息用[压缩](https://zh.wikipedia.org/wiki/%E6%95%B0%E6%8D%AE%E5%8E%8B%E7%BC%A9)形式表示[数字](https://zh.wikipedia.org/wiki/%E6%95%B0%E5%AD%97)[语音](https://zh.wikipedia.org/wiki/%E8%AF%AD%E9%9F%B3)[信号](https://zh.wikipedia.org/wiki/%E4%BF%A1%E5%8F%B7)[谱包络](https://zh.wikipedia.org/w/index.php?title=%E8%B0%B1%E5%8C%85%E7%BB%9C&action=edit&redlink=1)。它是最有效的音频分析技术之一，也是低位速下编码方法高质量语音最有用的方法之一，它能够提供非常精确的语音参数预测。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。

### 1.2.2 卷积神经网络研究现状

卷积神经网络（Convolutional Neural Network, CNN）为深度学习中一种重要的网络架构。它是一种[前馈神经网络](https://zh.wikipedia.org/wiki/%E5%89%8D%E9%A6%88%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C)，它的人工神经元可以响应一部分覆盖范围内的周围单元。卷积神经网络由一个或多个卷积层和顶端的全连通层（对应经典的神经网络）组成，同时也包括关联权重和[池化](https://zh.wikipedia.org/w/index.php?title=%E6%B1%A0%E5%8C%96&action=edit&redlink=1)层（pooling layer）。这一结构使得卷积神经网络能够利用输入数据的二维结构。

早在上个世纪六十年代，（Wiesel，Hubel，1965）通过对猫视觉皮层细胞的研究，提出了感受野（receptive field）的概念，这标志着神经网络结构首次在大脑视觉系统中被发现。而日本学者（Fukushima等，1980）在Wiesel与Hubel感受野概念的基础上提出了神经认知机(neocognitron)。神经认知机可以看作是卷积神经网络的第一个实现，也是感受野概念在在人工神经网络领域的首次应用。神经认知机将一个视觉模式分解成许多子模式（特征），然后进入分层递阶式相连的特征平面进行处理，它试图将视觉系统模型化，使其能够在即使物体有位移或轻微变形的时候，也能完成识别。通常神经认知机包含两类神经元，即承担特征抽取的 S-元和抗变形的 C-元。S-元中涉及两个重要参数，即感受野与阈值参数，前者确定输入连接的数目，后者则控制对特征子模式的反应程度。许多学者一直致力于提高神经认知机的性能的研究：在传统的神经认知机中，每个S-元的感光区中由 C-元带来的视觉模糊量呈正态分布。如果感光区的边缘所产生的模糊效果要比中央来得大，S-元将会接受这种非正态模糊所导致的更大的变形容忍性。我们希望得到的是，训练模式与变形刺激模式在感受野的边缘与其中心所产生的效果之间的差异变得越来越大。为了有效地形成这种非正态模糊，Fukushima 提出了带双 C-元层的改进型神经认知机。

卷积神经网络结构的确立源自于1997年Yann LeCun的一篇论文（Yann Le Cun等，1997），他们设计了一种名为LeNet-5的多层人工神经网络，可以对手写识别数字做分类。卷积神经网络也像其他神经网络一样可以使用反向传播算法进行训练。在当时的技术条件下，LeNet-5就可以实现低于1%的错误率。由此该算法结构得到了当时整个美国绝大多数邮政系统的应用。LeNet-5可以算作是第一个产生实际商业价值的卷积神经网络应用。

直到2006年，由Geoffrey Hinton等人提出的深度置信网络（Hinton等，2006）与受限玻尔兹曼机（Salakhutdinov等，2007）的学习算法才重新使人工智能领域对神经网络产生了足够的关注。卷积神经网络的热潮的掀起则是由于2012年开始举办的ImageNet图像分类比赛。赛中（Krizhevsky等，2012）提出了一个经典的卷积神经网络结构，并在图像识别任务上取得了重大突破。其方法的整体框架叫做 AlexNet，与 LeNet-5类似，但层次结构上要更加深一些。同时使用了非线性激活函数ReLu与Dropout方法，取得了卓越的效果。

### 1.2.3 音频场景识别研究现状

Sawhney和Maes在1997年MIT媒体实验室的技术报告中提出一种专门解决ASC问题的方法。作者记录了一组包括“人”，“声音”，“地铁”，“交通”和“其他”的一组数据集。他们利用语音分析和听觉研究借鉴的工具从音频数据中提取了几个特征，采用递归神经网络和k最近邻标准对特征和类别之间的映射进行建模，并获得68％的整体分类准确率。一年后，来自同一机构的研究人员Clarkson等通过戴着麦克风录制连续的音频流，同时进行一些超市自行车旅行，然后自动将音频分割成不同的场景（如“家”，“街道”和“超市”）。他们将从音频流中提取的特征的经验分布拟合成隐马尔可夫模型（HMM）。

与此同时，实验心理学的研究则着重于理解驱动人类对声音和场景进行分类和识别的能力的感知过程。 Ballas发现识别声音事件的速度和准确性与刺激的声学性质、它们发生的频率及是否它们可以与物理原因或声音刻板印象相关联有关（Ballas，1993）。佩尔顿等人（Peltonen等，2001）观察到人类对音频场景的认识是通过识别典型声音事件（如人声或汽车发动机噪声）来实现的，并且确定了人类识别25个声场中的能力的整体准确率为70％。 Dubois等人（Dubois等，2006）研究了在不是实验者先验的情况下，个体如何定义他们自己的语义类别分类。最后，Tardieu等人（Tardieu等，2008）测试了语义类的出现以及在火车站范围内对声场的识别。他们在报告中说，声源、人类活动以及房间效应（如混响）是促成音频场景形成的因素，也是类别为固定先验情况下的识别线索。

受心理声学/心理学文献的影响，这些文献强调音频场景识别的局部特征和全局特征，一些麻省理工学院研究人员则侧重于音频的时域特征。 Eronen等人（Eronen等，2003）采用Mel频率倒谱系数（MFCCs）来描述音频信号的局部频谱包络，用高斯混合模型（GMMs）来描述其统计分布。然后，他们通过利用训练信号种类的知识的判别式算法来训练HMM，以解释GMM的时域演变。 Eronen及其合作者通过考虑更多的特征，和在分类算法中增加一个特征变换步骤，进一步推进了这项工作，在18种不同的声场中获得了总体58％的准确性。

尽管关于ASC系统的文献丰富，但研究界缺乏协调一致的标准来评估和测试解决这个问题的算法。2013年，IEEE音频和声学信号处理（AASP）技术委员会首次组织了DCASE (音频场景和事件检测和分类)挑战赛，以测试和比较ASC和事件检测与分类算法。这一举措符合信号处理领域旨在促进可再生研究的目标。过去几年来，本挑战赛中已经提出了许多音频处理技术，对整个ASC系统的发展做出了极大的贡献。

Santoso等人（Santoso等，2016）使用卷积神经网络网络来构建分类器。该系统是基于计算机视觉领域工作中采用的体系结构而设计的，具体来说，用来构建本系统分类器的结构为NIN（网络中的网络）（Lin等，2013）。NIN体系结构被提出来改善局部模型在CNN卷积层的抽象能力。 NIN用一个更有效的非线性逼近器替代了对CNN中的数据补丁进行抽象的模型。 在NIN架构中，抽象模型被MLP网络取代。此外，NIN架构取代了传统CNN的分类方法。 在CNN架构中，特征映射被连接到作为分类器的传统MLP网络。 NIN架构使用全球平均汇集来取代这种分类方法。 NIN架构直接使用最后一个卷积层中的特征映射来构建分类器。 该体系结构取得特征映射的平均值，并将生成的向量直接输入到softmax层。在本系统的特征提取部分，使用梅尔频谱系数（MFCC）作为分类器的输入向量。分类器使用来自MFCC特征集的每个帧进行训练，然后对每个帧的结果进行阈值化并投票选择音频数据的最终场景标签。其系统准确度胜过DCASE挑战的基准系统。系统平均准确率为78.83％，基准系统平均准确率为72.57％。

辛德勒等人（Schindler，2016）通过使用CQT（常数Q变换）特征作为CNN的输入来增强结果。CQT是一个时频表示，其所有频段的Q因数,即中心频率与带宽的比值相等。 CQT本质上是一种小波变换，这意味着对于低频率，频率分辨率更好，时间分辨率对于高频率更好。使用CQT的动机来自于音乐感知领域的观点：人类听觉系统在大部分可听频率范围内近似为“常量Q”。该系统的关键是利用CQT以足够的分辨率捕获来自低频和高频的基本音频信息，并创建一个并行CNN架构，该架构能够及时捕获这两种频率。所呈现的深度神经网络架构已经比DCASE 2016声场景分类任务组织提供的基线系统超出10.7％的相对改进，在开发集合上达到80.25％。此外，它在验证集中达到了83.3％。

由于此前在音频场景分类领域缺乏大型标记的声音数据集，获得这些数据集通常既昂贵又不明确。来自MIT的Aytar等人寄希望于通过利用视觉和声音之间的自然同步来学习来自未标记视频的音频特征来扩大规模，因此他们利用超过一年的野外采集的声音来学习语义丰富的音频特征（Aytar等，2016）。未标记的视频可以大规模、低成本的获得，且具有音频信号。计算机视觉方面的最新进展使机器能够高精度地识别图像和视频中的场景和对象。而如何将视频中的知识转化为标记音频的标签成为了研究的关键。在实验中，他们使用了可以直接在原始音频波形上学习的卷积神经网络，通过将知识从视觉传输到声音进行训练。尽管网络是通过视觉监督进行训练的，但网络在推理过程中不依赖视觉。结果表明，与简单的全连接的网络或较早的图像分类体系结构相比，最先进的图像网络在音频分类方面具有出色的结果。其对较大的标签集词汇进行训练可以提高性能，尽管在对较小的标签集进行评估时性能稍有提高。

## 1.3 研究内容

通过之前的总结我们可以发现，此前应用的一般音频场景识别方法如GMM，HMM等存在着对音频集利用不充分的问题。根据之前的经验，本文采用卷积神经网络作为分类器，以提高音频集的利用效率。

具体研究内容章节如下：

第二章为音频场景分类基线系统的介绍与实验。主要介绍了特征提取的关键特征MFCC原理及其提取过程。并且对GMM分类器进行了详细介绍，主要包括了EM算法的原理与其在GMM模型中的应用。最后，进行了基线系统的测试，得到了基线系统的分类准确率。

第三章为基于卷机神经网络的音频场景分类系统的设计与实现。首先介绍系统的总体架构，然后详细介绍卷积神经网络的关键结构、原理及特点，再根据这些基础知识探讨将卷积神经网络应用在音频场景分类领域的适用性。接着，将探讨音频场景分类中卷积神经网络的训练方法。最后，根据之前的设计对模型进行训练并得出结论。

第四章为在前一章初步设计好的系统上进行一定的优化。本章中在原来系统的基础上，通过引入新的特征，来增强分类准确率。最后在实验中将与第二章的基于GMM分类器的模型进行对比，分析本文提出系统的优缺点。

# 第2章 音频场景分类基线系统

## 2.1 系统结构简介

本章主要内容为介绍基于MFCC特征及GMM分类器的音频场景分类基线系统。MFCC作为语音识别领域的重要特征，有着良好的区分性能。而且研究也表明（Reynolds等，1995），在GMM作为分类器与MFCC等性能良好的特征配合时，有着出色的分类性能。因此选取以上经典方法作为音频场景分类的基线系统，以判断之后提出的基于卷积神经网络的分类器是否具有优势。

本文中的基线系统主要包括五个部分，分别为音频数据集、特征提取与处理、系统训练、系统测试、系统评估。流程图如图2-1所示。



#### 图2-1 基线系统结构

本章首先介绍音频特征的提取与处理部分，然后讨论GMM模型原理及其核心算法，最后，介绍实验的环境、数据集、评价指标与结果。

## 2.2 特征的提取与处理

音频可以用多种方式表示，哪一种“最佳”取决于应用以及处理机器。多年来，特征设计和选择是许多音频分析任务的关键组成部分，经常使用的特征包括频谱形状的频谱质心和高阶统计，零交叉统计，谐波，基频和时间包络。

在分类任务，尤其是音频分类任务中，描述光谱形状的梅尔频率倒谱系数（MFCC）具有悠久的历史。尽管MFCC提取过程中会造成数据的有损压缩，但其分类与识别效果在数据速率很低时也具备相当的可用性，因此得到了广泛的应用。本小节将详细介绍MFCC特征的提取过程。

### 2.2.1 MFCC特征

提取音频信号的最佳参数表示是产生更好识别性能的重要任务。这一阶段的效率对下一阶段很重要，因为它会影响之后的分类效率。

人类之所以能在复杂的声音环境中判断出不同的环境，主要在于耳蜗的功劳。耳蜗可以看作为一个滤波器组，帮助人们过滤20-20KHz的音频。问题在于耳蜗对于听觉范围内频率的灵敏度并不是线性的，存在一种映射关系。因此为了模拟出人耳的频率响应，（Davis等，1980）提出了MFCC特征。

MFCC的整个过程如图2-2所示：

****

#### 图2-2 MFCC步骤

如图2所示，MFCC由七个步骤组成。下面将具体介绍这些步骤：

（1）预加重

许多现实中的音频信号，它们的功率谱随频率的增加而减小，其大部分能量集中在低频范围内。这就造成了音频信号中高频的信噪比可能降到不能容许的程度。预加重的目的是更加凸显音频信号的高频特性。通常借助高通数字滤波器来实现。

设第n时刻的语音采样值为x(n)，经过预加重处理后的结果是:  
其中a为预加重系数，通常取0.9~1.0之间。

（2）分帧

分帧将从模数转换（ADC）中得到的音频样本分割为长度在20至40毫秒范围内的小帧。音频信号被分N个样本的帧。相邻帧由M（M<N）分隔。通常取M=100，N=256。

（3）加窗（海明窗）

在预加重与分帧完成后，需要为每一帧加上海明窗（Hamming window）。加窗是为了控制数据处理量，每次仅处理窗中数据。由于之后FFT中的处理对象是有限长信号，对无限长信号强行进行有限个点的FFT会丢失频率信息导致频谱泄露。海明窗的公式可以总结如下：

假设窗函数为：。且为每帧中的样本数，为输出信号，为输入信号，则加窗结果如下：

其中，

（4）快速傅立叶变换（FFT）

将N个样本的每个帧从时域转换到频域。具体过程如下：

（5）Mel滤波

FFT频谱中的频率范围非常宽，语音信号不遵循线性标度。因此通过如图2-3所示的Mel标度滤波器组。



#### 图2-3 Mel标度滤波器组

该图表示了一组三角形滤波器，用于计算滤波器频谱分量的加权和，使得处理后的输出近似于Mel标度。每个滤波器的幅度频率响应是三角形的，在中心频率处等于1，在两个相邻滤波器的中心频率处线性减小到零，且每个滤波器输出是其滤波后的频谱分量的总和。最后，使用以下等式计算给定频率f的Mel：

（6）离散余弦变换

该过程为使用离散余弦变换（DCT）将log-Mel频谱转换为时域的过程。转换的结果称为Mel频率倒谱系数。系数集称为声矢量。因此，每个输入都被转换为声学矢量序列。

（7）差分计算

以上取得的倒谱参数只能反应音频信号的静态特性。为了提高信号的识别性能，应采用音频信号静态特性的差分谱来描述音频信号的动态特性。因此引入了13个一阶差分特征（12个倒谱特征加能量），以及39个二阶差分特征。从时间t1到t2的窗口中信号x的帧能量如下式所述：

13个一阶差分特征中的每一个特征表示MFCC特征中对应的倒谱或能量特征的帧之间的变化，而39个二阶差分特征中的每一个表示对应的一阶差分特征中的帧之间的变化。其中一阶差分的计算如下式所述：

其中，表示时刻的倒谱系数。

## 2.3 GMM模型原理

GMM（Gaussian Mixture Model）即高斯混合模型，是表示为高斯分量密度加权和的参数概率密度函数。目前GMM已经广泛应用于语音识别、音频场景识别、音频检索等领域，得到了研究人员的广泛的认可。GMM使用迭代期望最大化（EM）算法从训练数据估计中GMM参数。下面将详细介绍GMM模型。

### 2.3.1 高斯混合模型

M阶GMM是由M个高斯概率密度函数加权求和得到的，如下式所述：

其中，是D维连续特征向量，，为加权系数，为高斯密度分量。每一个密度分量可表示如下：

其中是均值向量，是协方差矩阵，且M个权重满足约束条件。完整的高斯混合模型由均值向量，协方差矩阵和混合权重组成，统一表示如下：

利用GMM进行音频场景分类的核心思想在于：在音频文件经过特征提取与处理后，我们可以得到特征序列。在忽略时序信息的条件下，通过以帧为单位，利用GMM对音频信息进行建模。由于音频事件中每帧信号对应的特征划分为若干类，类与类之间的音频特征又相互独立，且音频特征均服从相同的正态分布。所以可以将多个类的正态分布按照一定权重进行组合，代表某类音频特征的总体分布，以表现该音频场景的特征。

### 2.3.2 GMM的参数估计

在给定了训练向量和GMM配置后，我们希望估计GMM的参数，该参数在某种程度上与训练特征向量的分布最匹配。有几种技术可用于估计GMM的参数，迄今为止最流行和最成熟的方法是最大似然估计。

最大似然估计的目的是在已知训练特征向量集的情况下，找到使GMM似然函数最大的模型参数。对一组长度为T的训练特征向量序列，假设向量间相互独立，则GMM似然度可表示为：

但上式为参数的非线性函数，无法直接求出极大值，故不能显式的求出参数。然而，可以使用期望最大化（EM）算法的特殊情况迭代地求出最大似然参数估计。

### 2.3.3 EM算法

EM算法（Dempster等，1977）是当数据不完整或有缺失值时从给定数据集中找到基础分布的参数的最大似然估计的一般方法。EM算法的每次迭代由两部分组成，第一步为E步，即求期望（expectation）过程；第二步为M步，即求极大（maximization）过程。故该算法简称为EM算法。由上一小节总结可以得出，在概率模型中含有隐变量时，不能依赖于极大似然估计法，进而引入EM算法以解决含有隐变量时概率模型参数的极大似然估计。那么EM算法是如何保证在每一次的迭代中都使得模型的似然度增大呢？

我们假设为一组长度为T的训练特征向量序列，模型参数的似然度为，则GMM的训练目的在于找出使最大的模型参数，也即：

由于为参数的非线性函数，无法直接求最大值，故引入Q函数（Nuttall等，1972），应用如下：

其中，为高斯分量序号，为估计的新模型参数，为在模型参数为时，归于高斯分量序号为的概率密度。

进一步的，有：

故：

当且仅当时，等号成立。因为与具有相同的单调性，所以只需证明是关于对数函数的凹函数函数即可。对取关于的微分：

将带入上式，可以得到：

当时，与极值点相同。故与有相同的单调性和极值点。由此可以得出结论：可以通过求的局部极大值来得到的新模型参数。

### 2.3.4 EM算法在GMM模型中的应用

如前所述，GMM算法的输入数据为一组长度为T的训练特征向量序列。要求的结果为，即新模型参数。在每次EM迭代中，首先设定每个参数的初始值。接着，根据当前的模型参数，计算分模型对的响应度:

此为EM算法中的E步。然后，使用以下重估公式计算新一轮模型的迭代参数：

混合权重：

均值向量：

方差（对角协方差）：

此为EM算法中的M步。

最后，重复E步与M步，直至收敛。

## 2.4 实验准备与结果

### 2.4.1 实验环境与系统细节

本实验在以Linux为内核的Ubuntu系统下实现。实验使用Python语言，引入了外部库Librosa进行特征提取。在训练部分引入了sklearn库，以进行应用GMM的非监督学习。音频特征部分包括60维MFCC特征向量，其中包括了20个MFCC静态系数（包括第0个）、 20个一阶差分系数、20个二阶差分系数。分帧过程的帧长为40ms，帧移为帧长的50%即20ms，加窗过程采用汉明窗。对于每个音频场景，根据特征标签使用EM算法训练具有16个分量的GMM模型。测试阶段使用最大似然判定法判断音频场景的归类。使用准确度来衡量分类性能。

系统的实现细节如表2-1所示：

#### 表2-1 系统实现细节

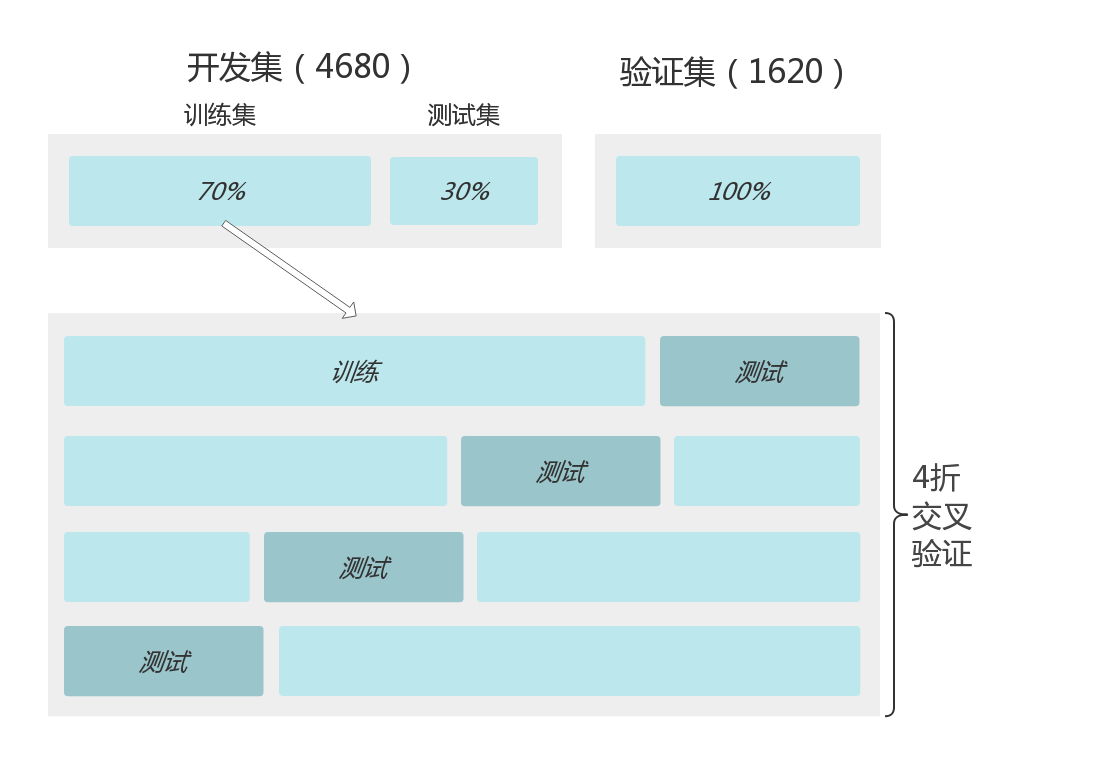
|  |  |
| --- | --- |
| **音频输入** | |
| 通道 | 单通道（多通道平均至单通道） |
| 标准化处理 | 无 |
| **音频特征** | |
| 类型 | MFCC |
| 窗口长度 | 40ms |
| 跳跃长度 | 20ms |
| **特征向量** | |
| 构成 | 静态MFCC+一阶+二阶 |
| 向量长度 | 60 |
| **高斯混合模型** | |
| 高斯分量数 | 16 |
| 协方差 | 对角线 |
| 参数数量 | 1936 |
| 模型 | 每个场景对应一个模型 |
| 决策方法 | 似然积累+最大值 |
| 积累窗口 | 信号长度 |

### 2.4.2 实验数据集

本实验使用的数据集为TUT Acoustic Sc​​enes 2017，负责采集的团队为坦佩雷理工大学音频研究小组。该数据集由15个不同标签的音频场景组成：湖畔海滩，公共汽车，咖啡馆/餐厅，汽车，市中心，森林小径，杂货店，家庭，图书馆，地铁站，办公室，城市公园，住宅区，火车和电车。每个场景包含312个音频文件，所有音频文件都被切割成长度为30秒的片段。为了满足所有音频场景类别的高声学可变性的要求，每次录音均在不同的位置进行，录制的平均持续时间为3-5分钟。用于记录此特定数据集的设备包括双声道Soundman OKM II Klassik / studio A3入耳式麦克风和使用44.1 kHz采样率、24位分辨率的Roland Edirol R09波形记录仪。

数据集分为开发集和验证集两部分，其中，开发集包含4680个音频文件，其中大约70%的数据用于训练音频场景分类模型，剩下30%用作测试。可以从数据集中提供的元数据文件或音频文件名中找到类别标识符。

而在验证集中，共有1620个音频文件，其中每类音频文件108个。每段音频长度为10秒，每个场景音频总计18分钟。验证集中的文件皆为重新采集的音频，且全部数据都用于验证系统性能。验证集的所有文件在测试时都不带有标签，只会在测试结束后才会从独立的标签文件中提取音频的对应标签，因此验证集的分类准确率可能会显著低于开发集。



#### 图2-4 实验数据集详解

由于样本量较小，在训练阶段为了充分利用数据，将数据集等分为四份，其中每一份中各类型音频比例与总数据集一致。任意取其中一份作为测试集验证训练结果，其他三份用做训练数据，然后换其中另一份作测试集，其他份用于训练，如此操作共四次直到每份数据都被用于测试。此种方法称为4折交叉验证法（4-fold Cross Validation），保证了可以生成四组不同的训练与测试集。

### 2.4.3 评价指标

在使用训练集对GMM模型进行训练后，我们通过测试集来测试模型训练结果。训练结果的唯一评判标准为准确率（ACC），即预测正确的样本数与总样本数间的比值。计算公式如下：

其中，为预测正确的样本数，为总样本数。

### 2.4.4 实验结果与分析

如表2-2所示，通过4折交叉验证法的训练与测试，一共得出了四组准确率数据，且因为数据在切分时的总量与各类别比例都相同，因此总准确率为四组准确率的平均值。从数据中我们可以看出，不同子数据集之间因数据差异，训练准确度也各有差异，但整体准确度维持在了74%左右，可以评价该基线系统对于不同数据能实现相对稳定的分类水平。

#### 表2-2 不同子集的总分类准确率 （单位：%）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **场景** | **总体** | **1** | **2** | **3** | **4** |
| 沙滩 | 76.0 | 79.5 | 78.2 | 96.2 | 50.0 |
| 公交 | 83.0 | 89.7 | 75.6 | 87.2 | 79.5 |
| 咖啡/餐馆 | 78.5 | 61.5 | 93.6 | 84.6 | 74.4 |
| 汽车 | 92.0 | 88.5 | 89.7 | 89.7 | 100.0 |
| 市中心 | 86.5 | 78.2 | 94.9 | 76.9 | 96.2 |
| 森林小径 | 66.7 | 74.4 | 79.5 | 65.4 | 47.4 |
| 杂货铺 | 74.4 | 89.7 | 71.8 | 74.4 | 61.5 |
| 家 | 72.0 | 85.9 | 80.2 | 60.5 | 61.5 |
| 图书馆 | 60.6 | 16.7 | 73.1 | 75.6 | 76.9 |
| 地铁站 | 77.2 | 83.3 | 71.8 | 69.2 | 84.6 |
| 办公室 | 97.8 | 96.2 | 100.0 | 96.2 | 98.7 |
| 公园 | 48.4 | 75.6 | 44.9 | 21.8 | 51.3 |
| 小区 | 73.7 | 70.5 | 70.5 | 78.2 | 75.6 |
| 火车 | 36.2 | 38.5 | 29.5 | 20.5 | 56.4 |
| 电车 | 81.7 | 91.0 | 62.8 | 88.5 | 84.6 |
| **平均** | **73.6** | **74.6** | **74.4** | **72.3** | **73.2** |

如表2-3所示，在验证集上重新运行模型后，整体分类性能相比开发集有了明显下降。分析其原因，就在于使用验证集测试时不带有测试标签，而之前训练过程中存在对特定数据的记忆效应，在验证阶段新数据带来的随机性会导致分类性能的下降。由于验证集的测试数据量相比开发集更大，因此更能反映出系统的真实分类性能。

#### 表2-3 开发集与验证集分类准确率对比（单位：%）

|  |  |  |
| --- | --- | --- |
| **场景** | **开发集** | **验证集** |
| 沙滩 | 76.0 | 30.6 |
| 公交 | 83.0 | 41.7 |
| 咖啡/餐馆 | 78.5 | 61.1 |
| 汽车 | 92.0 | 57.4 |
| 市中心 | 86.5 | 79.6 |
| 森林小径 | 66.7 | 81.5 |
| 杂货铺 | 74.4 | 61.1 |
| 家 | 72.0 | 96.3 |
| 图书馆 | 60.6 | 23.1 |
| 地铁站 | 77.2 | 94.4 |
| 办公室 | 97.8 | 69.4 |
| 公园 | 48.4 | 15.7 |
| 小区 | 73.7 | 80.6 |
| 火车 | 36.2 | 63.0 |
| 电车 | 81.7 | 55.6 |
| **平均** | **73.6** | **60.7** |

## 2.5 本章小结

本章首先介绍了音频场景分类基线系统的系统结构，然后给出了MFCC特征的提取过程；再以EM算法为核心，介绍了GMM模型的原理及EM算法在GMM模型中的应用；最后，进行了基线系统的实验。实验包括了实验环境、实验数据集介绍，系统性能评价指标以及实验结果和分析。该系统在训练集上实现的分类准确率为73.6%，验证集上分类准确率为60.7%。

# 第3章 基于卷积神经网络的音频场景识别系统

## 3.1 引言

卷积神经网络（Convolution Neural Network，简称CNN）目前已广泛用于语音识别，计算机视觉和自然语言处理应用。这些技术最近在各种应用的音频分类领域取得了巨大成功。尽管之前曾将卷积神经网络主要用于视觉识别中，但卷积体系结构也已成功应用于语音和音乐分析。

得益于卷积神经网络结构能够从高维原始数据中学习分层特征,使得卷积模型优于基于手工设计特征的常用方法，并达到与其他特征学习方法类似的水平。虽然训练时间可能会更长且结果远非突破性，但卷积神经网络可以有效地利用现有的音频数据集，即使数据集的数量可能非常有限且数据难以获取。更重要的是，可用数据集大小的显着增加很可能极大地改善训练模型的性能。

由于此前经过验证，MFCC作为特征具有良好的分类性能，因此在本章系统中仍然使用其作为音频特征中的重要特征。系统流程设计如图3-1。



#### 图3-1 音频分类系统的基本流程

## 3.2 卷积神经网络原理

卷积神经网络作为一种特殊的深层前馈神经网络，具有局部连接、权值共享、子采样的结构特点。得益于这些特点，使卷积神经网络剧有了缩放不变性、平移不变性及旋转不变性，同时也使卷积神经网络在计算时相比一般的前馈神经网络需要更少的参数，大大提高了效率。

一般的卷积神经网络主要由卷积层、激活函数、池化层与全连接层等组成。作为将卷积神经网络应用到音频场景分类领域的基础知识，本小节将分别介绍卷积神经网络中的重要结构、原理及特点。

### 3.2.1 卷积层

卷积层是卷积神经网络的核心模块，可以完成大部分繁重的计算工作。其作用为提取一个区域的局部特征。卷积层执行的核心操作称作卷积（convolution），卷积为分析数学中的常用运算方式，是通过两个函数生成第三个函数的一种数学算子。在机器学习领域的应用中，卷积通常体现为在一幅图像或某种特征上滑动一个滤波器，借助这样的操作以得到一组新的特征。其中二维离散序列卷积的定义如下：

假设有图像与滤波器，且，，则其卷积可以表示为：

如图3-2为一个二维卷积的示例：



#### 图3-2 二维卷积的示例

卷积运算的结果，如上图中等式右侧所示，称为特征图（feature map）。特征图为一幅图像经过卷积运算后提取到的特征。通常为了提升卷积神经网络的特征提取能力，可以在神经网络的每层使用多个不同特征。

而应用于本文音频特征的图像处理中，为了更加充分的提取出图像的局部信息，使用三维结构的神经层，其尺寸为高度宽度深度，即个的特征图组成。此外，还需要指定的超参数有滤波器的个数，滤波器的大小，步长以及边界填充。

生成的特征图大小的计算方法如下：

### 3.2.2 激活函数

一般在卷积层进行卷积运算之后，为了增强网络的表达与学习能力，会在卷积层之后加入连续非线性的激活函数（activation function）。原因在于连续非线性的激活函数可导，可以用最优化的方式来学习网络参数。

本文中使用的函数为ReLU（Rectified Linear Unit，线性修正单元），为当前卷积神经网络中常用的函数。其定义为：

ReLU本质上为一种斜坡函数，优点在于计算时只用进行加法、乘法与比较操作，无梯度耗散问题，收敛快，计算上更加高效。不同于已经濒临淘汰的sigmoid函数，当时，神经元的输出为0，增加了网络的稀疏性，使50%的神经元处于激活状态。

### 3.2.3 池化层

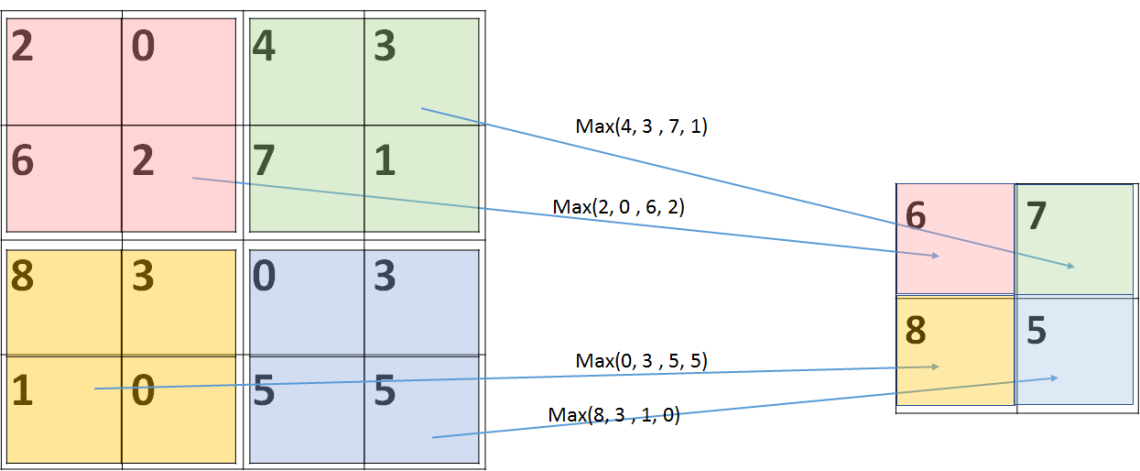
池化层（pooling layer），通常置于卷积层、激活函数之后，其对输入的特征图进行压缩，以减小图片尺寸进而简化网络复杂度；另一方面其进行特征选择，通过降低特征的数量以减少网络参数数量。

假设池化层输入的特征图组为，对于其中的每一个特征图，可以将其划分为子区域 。则池化的常见两种定义如下：

（1）最大池化（maximum pooling）

最大池化为在一个区域内寻找所有神经元的最大值，表述如下：

其中，为子区域中的输入特征图组，为经过最大池化的输出特征图组。最大池化的例子如图3-3所示：



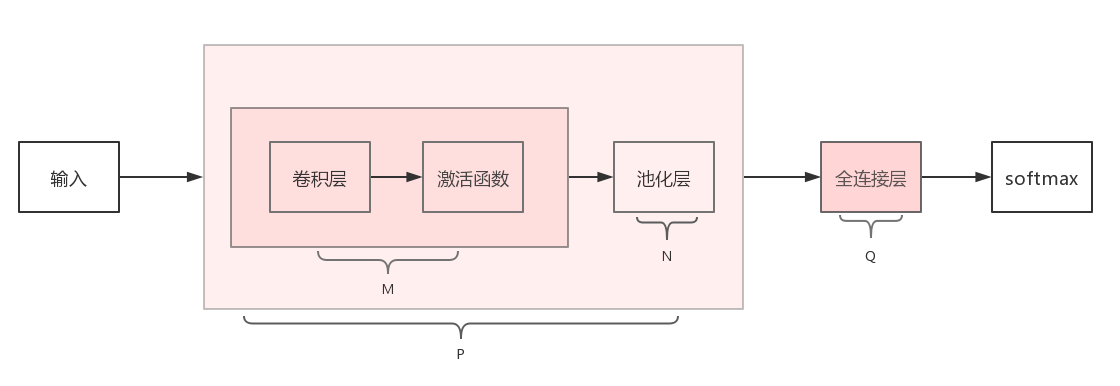
#### 图3-3 最大池化过程示例

（2）平均池化（mean pooling）

平均池化为在一个区域内取所有神经元的平均值，表述如下：

### 3.2.4 卷积神经网络的结构

典型的卷积神经网络通常由卷积层、池化层、激活函数与全连接层交叉堆叠而构成。结构如图3-4所示，通常卷积层与激活函数组合的个数M取2～5，池化层个数N取0或1。在经过P个（P取1～100）连续的卷积模块后，再通过Q个（Q取0～2）全连接层将所有特征连接，将输出值给softmax分类器。其中，由卷积层、池化层、激活函数与全连接层所构成的部分也成为隐藏层。



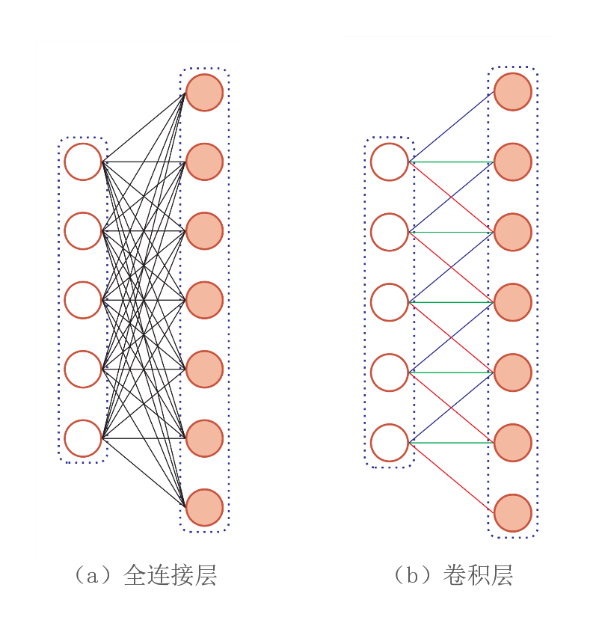
#### 图3-4 典型的卷积神经网络结构

### 3.2.5 卷积神经网络的特点

不同于全连接前馈神经网络，卷积神经网络的权重矩阵参数非常少，带来的结果是训练效率的大大提升。而区别的核心就在于卷积神经网络局部连接和权重共享的特点。

1. 局部连接

如图3-5(a)所示，在全连接层中，第层的每一个神经元都与下一层（即第*l*+1层）中的每一个神经元相连。故总连接数为。而在卷积层中，如图3-5(b)所示，每一个神经元都只与下一层中滤波器窗口内的神经元相连，构成了局部连接网络。连接数变为了（为滤波器尺寸，且通常），使连接数大大减小，进而提升了计算效率。



#### 图3-5 全连接层与卷积层

1. 权重共享

在卷积层的卷积运算中，第层的净输入可表示为：

其中，为权重向量，为第*l*-1层的激活值，为偏置。

参考公式3-8我们可以看出，滤波器权重作为参数对于第层的所有神经元都是相同的。由于同一特征映射面上的神经元权值相同，所以网络可以并行学习，这也是卷积网络相对于神经元彼此相连网络的一大优势。权重共享降低了网络的复杂性，特别是多维输入向量的图像可以直接输入网络这一特点避免了特征提取和分类过程中数据重建的复杂度。

由于局部连接与权重共享，导致了卷积层中的参数个数与层中神经元的数量没有关系，而只与滤波器尺寸和1维偏置有关，共计+1个参数。因此神经元的增多并不会严重影响卷积层的计算效率。

### 3.2.6 卷积神经网络的参数学习

卷积神经网络中网络参数的训练与全连接前馈网络类似，采用误差反向传播算法。但区别于全连接前馈网络的通过计算每一层误差项进行反向传播来计算每层参数梯度，卷积神经网络中的参数只有卷积核与偏置，故只需计算卷积层中参数即可。

假设第*l*层为卷积层，其前一层*l*-1层的输入特征图为。则经过*l*-1层卷积运算后，第*l*层的特征图净输入为。其中，第*l*层的第k个（）特征图的净输入为：

其中，为滤波器权重，为偏置。第*l*层共有卷积核个与偏置*K*个，其梯度可以使用链式法则计算。

损失函数关于第*l*层的滤波器权重的偏导数为：

其中，为损失函数关于第*l*层的第k个特征图净输入的偏导数。

同理，损失函数关于第*l*层第k个偏置的偏导数为：

由此可知，每层参数的梯度计算依赖于其所在层的误差项。而在卷积层和池化层中，误差项的计算又有区别，因此需要分别计算。

（1）卷积层的误差项计算

假设第*l+*1层为卷积层，特征图净输入，类似于式3-9，第*l*层的第k个（）特征图的净输入可表示为：

则第*l*层的第d个特征的误差项可借助式3-13推导：

其中，为第l第层激活函数的导数，为宽卷积。

（2）池化层的误差项计算

由于池化层进行的是下采样（subsampled）操作，故当*l*+1层为池化层时，第*l*+1层神经元的误差项对应于其上一层特征图的一个区域。根据连式法则，要求得第*l*层的一个特征图对应的误差项，只需将其下一层对应的特征图误差项进行上采样（upsampling）操作，再与第*l*层特征图激活值的偏导数逐元素相乘即可。

第*l*层的特征图误差项的推导如下：

其中，**up**为上采样函数。

## 3.3 卷积神经网络在音频场景分类中的适用性

此前卷积神经网络已成功用于各种音频相关任务，如语音识别，音乐分析、事件检测。而在音频场景分类中使用卷积神经网络作为分类器有以下几点原因：

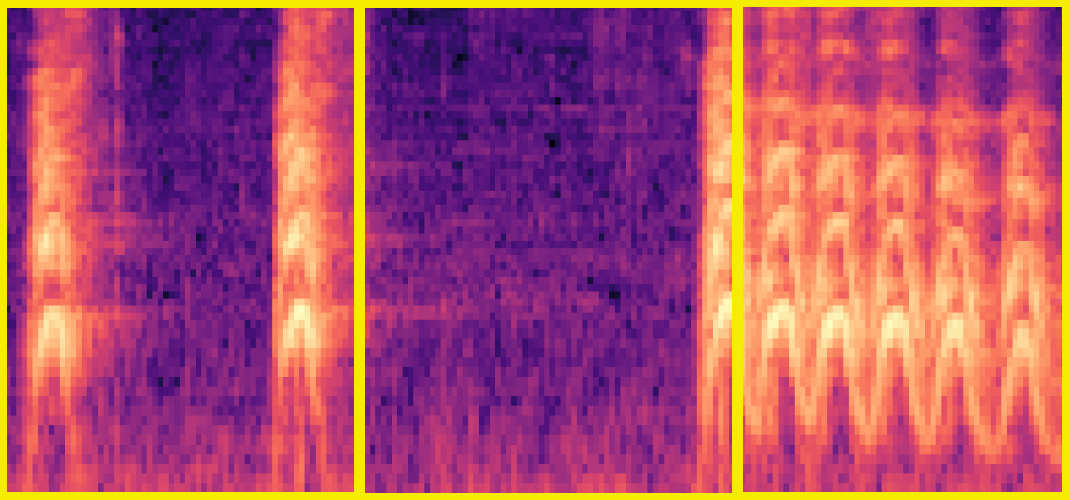
1. 卷积神经网络可以直接处理时频联合数据；
2. 卷积神经网络可以用自动学习的特征代替手动设计的特征进行分类，使分类更为高效；
3. 卷积神经网络具有良好的捕捉周期性时频特征的能力。

本小节将从卷积神经网络的各层出发研究卷积神经网络在音频场景分类中应用的可行性以及针对音频场景分类这一领域的特殊处理。

### 3.3.1 输入数据的处理

在将卷积神经网络应用于计算机视觉任务时，通常将各个图像分成不同的颜色通道如红色、绿色和蓝色。然后，每个隐藏层可以访问每个颜色通道中相应的感受野。将卷积神经网络应用于音频场景分类时可以采用同样的思想。

如图3-5所示，输入图像为log-Mel谱图，即在第二章Mel频谱图的基础上采用对数频率坐标。



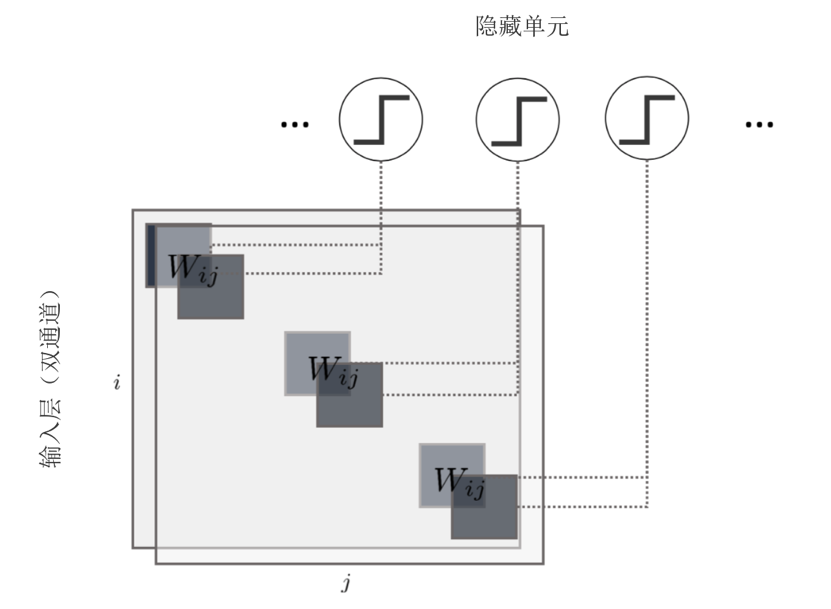
#### 图3-5 卷积神经网络的输入图像

此外，需要进一步操作以便卷积神经网络能输入固定大小的片段。如图3-5所示，log-Mel谱图被分割成更小的子片段（图3-5框中），每个子片段形成单独的输入数据。这样通过人为地增加了输入数据，提升了卷积神经网络的分辨率，同时降低了复杂性。

### 3.3.2 卷积层的应用

复杂声学场景包含有容易辨别的时频重复特征，如发动机噪音和电话铃声。这些特征称为局部模式，表现为频率和时间上能量的反复集中。例如，发动机噪声的特征在于跨越时间轴的局部模式，而铃声可以呈现跨越频率轴的重复特征。这种局部模式可以通过输入和一组滤波器权重之间的卷积运算来表示，它产生输出：

其中和是I的行和列索引，而*u*和*v*是滤波器权重的行和列索引。该卷积运算为离散二维卷积运算，在卷积层中进行。而卷积层对输入数据的应用如图3-6所示：



#### 图3-6 卷积层对输入数据的应用

其中：

（1）每个隐藏单元与输入层中坐标处的感受野之间为局部连接;

（2）输入的多个通道在感受野和隐藏单元之间保持相同的关系;

（3）使用相同的滤波器权重W，即构成权重共享，以捕获输入数据中的类似的重复特征。

### 3.3.3 池化层的应用

池化层置于每个卷积层的输出之后以降低其分辨率。最简单的池化操作是最大池化，其中池化层输入中的值块被替换为其单个最大值。应用于音频场景分类任务时，池化操作可将时频特征中的微小变化过滤掉。

例如，以特定频率为中心的相同局部模式（例如引擎噪声）可能只在一个记录与另一个记录间略微变化。池化过程允许降低频率或时间分辨率，将分类的重点转移到局部模式上。

### 3.3.4 全连接层的应用

卷积层和池化层可以按顺序复制，以增加深度并进而提取出更高级别的输入特征。在末端通过全连接层和softmax层实现分类。全连接层的输入为最后一层卷积层或池化层的输出。全连接层的作用为将输入分类为输出的音频场景之一。

每个输入片段（原始的全谱图）都是独立分类的，借助softmax函数识别最可能的音频场景，频谱图的分类根据多数投票进行。用于卷积神经网络训练的目标函数（即用于模型参数和的优化）为在N个输入样本上最小化目标和预测之间的损失函数：

## 3.4 系统设计

训练卷积神经网络需要做出很多关于架构（如输入数据的格式，卷积层的数量和大小，池化层的数量，滤波器维度）和学习超参数（学习率、动量、批次大小、丢失率）根据图3-1的系统设计流程图，本小节给出系统的设计细节。

由于培训完整模型所需的时间很长，详尽评估所有潜在组合是不可能的。 因此，综合效能最佳的模型的选择必须基于对最重要因素（层数/滤波器，滤波器形状，学习速率，丢失率）执行的有限验证（通常设置为10-20个迭代周期）。故设计阶段应考虑不同的训练参数作为对比，在有限的迭代周期内判定参数的适用性。而考虑中核心的要素即为分类平均准确率和单位迭代周期长度。最终得出的系统要在分类准确率尽可能高的同时训练时间在可控制范围内。

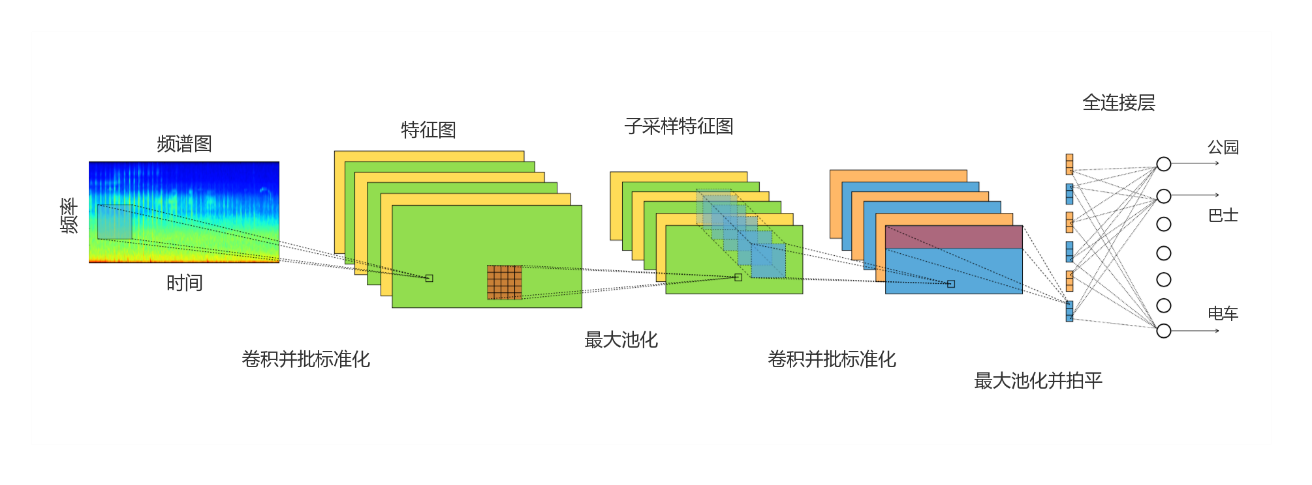
### 3.4.1 特征图与预处理

本系统选择的特征图为log-mel谱图。MFCC的优势在第二章已经阐述过，不再赘述。为了计算它，在40毫秒音频的窗口上应用短时傅立叶变换（STFT），并加50％的重叠和Hamming窗口。然后，对每个窗的绝对值进行平方，并应用60波段的梅尔滤波器组。最后，进行了梅尔能量的对数转换。使用librosa库在Python中实现了整个特征提取过程。

在提取过程之后，通过减去其平均值并除以其标准偏差来标准化每个窗，两者都在每个折叠的整个训练集上计算。 然后将归一化的谱图分成更短的谱图，下文中称之为序列。与用于STFT的帧不同，此处选择序列不重叠。在该过程结束时，卷积神经网络的输入是矩阵，可以将其视为单声道图像。

### 3.4.2 卷积神经网络架构

系统核心的卷积神经网络模型如图3-7所示：



#### 图3-7 卷积神经网络模块构成

第一层在输入频谱图上执行卷积。由于卷积滤波器的大小体现为局部感受野的大小，不同尺寸的卷积滤波器大小决定其提取特征的粒度，故实验中会安排不同尺寸的滤波器大小以进行对比。此外，滤波器的数量也会影响到特征的分析角度，因此实验过程中也会安排不同的滤波器数量进行对比。滤波器数量变多会增加特征分析的角度，但同时也会使计算量上升，而且过多的滤波器可能导致参数冗余。通常，滤波器的数量取。在卷积过程执行完成后，使用最大池化层对所获得的特征图进行子采样。

第二卷积层与第一卷积层相同，除了使用更多内核（取第一层的2倍）以便在更高级别上的表示特征。然后，针对时间轴的“破坏”执行第二次和最后一次子采样。因此，依然使用最大池化层，其在整个序列长度上操作。卷积层中用于内核的激活函数是ReLU。

最终，由于分类涉及10个不同的类，最后一层是由10个完全连接的神经元组成的softmax层。

音频的分类通过序列的预测分数求平均来进行。卷积神经网络的输出是一个包含第*i*个序列的所有分类预测分数的向量，整个段的预测类*c \**计算如下：

（待润色）其中，M是音频的种类数，为。 换句话说，预测的类是由每个序列输出的所有预测矢量的平均值给出的矢量中的最大条目yc \*的位置。

该系统使用Python语言编写的Keras库，Kears的后端基于Tensorflow库。用于训练的损失函数是分类交叉熵，为其最小化选择的优化算法是自适应动量，即Adam optimizer（Adam等，2014）。

### 3.4.3 批标准化

（Ioffe等，2015）中介绍的批标准化是一种解决（Shimoidara等，2000）描述的内部协变量转换问题的技术。我们可以将批标准化技术看作放置在两个卷积层中的每一个之后的中间层，以便白化这些层的输出。这种做法可以大大减少训练收敛时间。批量标准化层将线性变换应用于其输入，如下所示：

其中和是批标准化层输入的平均值和方差。此外，*γ*和*β*表示在训练期间将学习的变换参数。由于使用批量标准化来规范化内核输出，因此每个内核都有一个*γ*和一个*β*。

通过使用批标准化，虽然增加了模型的复杂性，但使整个系统获得了更好的性能并且大大减少了训练收敛所需的迭代周期数。

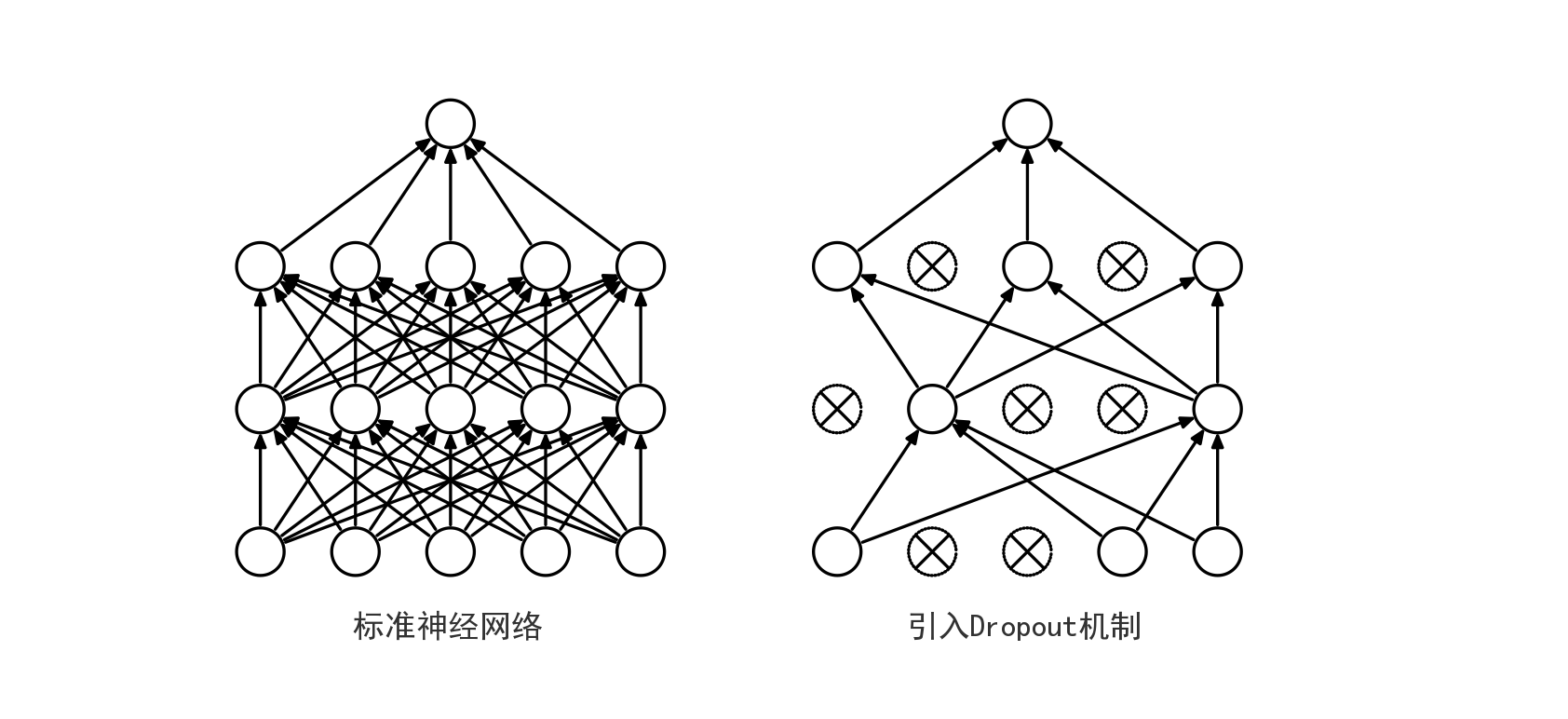
### 3.4.4 Dropout机制

在卷积神经网络模型中，如果参数过多且训练样本过少，经常会发生过拟合（overfitting）现象。过拟合具体体现为：模型在训练阶段损失函数较小，预测准确率较高，但在测试阶段损失函数较大，准确率偏低。

为了解决过拟合现象，通常使用模型集成方法，即多个模型组合训练。而模型集成带来的问题就是训练过程中成本过大，过于费时，且测试阶段同样要耗费大量时间。

（Hinton等，2012）在其论文中首次提出了Dropout机制，Dropout通过在训练过程中，通过忽略固定数量（通常为一半）的特征检测器，可以有效缓解过拟合的发生，使神经网络训练加速。

假设神经网络的输出为*x*，输出为*y*。则标准神经网络的流程为将*x*通过网络前向传播，之后将误差反向传播，更新参数让网络学习。而Dropout引入后，会首先随机隐藏网络中一半的隐藏神经元，如图3-8，而输入输出神经元保持不变。然后依然将输入*x*通过修改后的网络进行前向传播，把得到的损失结果通过网络进行反向传播。经过一批训练样本后，在没有删除的神经元上按梯度下降法更新参数。最后，继续重复该过程。



#### 图3-8 Dropout机制

由于随机删掉一半隐藏神经元导致网络结构发生变化，整个Dropdout过程相当于对多个不同的神经网络取平均。通过引入Dropout机制产生的平均作用，抵消了网络中部分相反的拟合，进而达到类似模型集成减少过拟合的效果。此外，由于Dropout过程可能会造成某两个神经元不会总是同时出现的情况，可以减少神经元之间复杂的共适应关系。这样一来，权重更新不会再依赖于网络中的某些固定结构，从而使整个网络的学习更具有健壮性。

### 3.4.5 模型训练

系统训练包括两个阶段。第一阶段称为非完全训练，首先将整个训练数据分成两个子集：一个用于训练，一个用于验证。每个迭代周期都会将训练频谱图收集到分类特征列表中，以便在序列分类之前随机改变它们并对其进行时移，这样做是为了增加输入可变性。然后，每隔固定迭代周期（本文取5）检查训练和验证集上的分段性能。检查后，如果分段验证分数得到改善，将保存网络参数。最后，如果在足够长的迭代周期之后（本文取15）没有记录到任何改进，会停止训练。通过这种方法，会使分段验证性能容易饱和。这意味着得分开始在固定的稳定值附近振荡。当发生这种情况时，说明系统已经收敛，因此可以进入第二阶段，即完全训练阶段。

在完全训练阶段，在固定数量的迭代周期的所有训练数据上重新训练网络。通过观察非完全训练期间分段验证准确度的收敛时间来选择该训练迭代周期。以这种方式训练的模型将达到收敛状态且不会过度拟合，并充分利用所有可用的训练数据。在处理小数据集时，训练效果可能变得特别理想。

## 3.5 实验准备与结果

### 3.5.1 实验环境与数据集处理

本章实验依然使用TUT Acoustic Scenes 2017 数据集，此处不再介绍。对于每个10秒的音频文件，使用40ms的分析帧和50％跳跃大小的窗口在40个频带中提取对数梅尔能量。神经网络由两个CNN层和一个完全连接的层组成，并使用大小为40x500的输入，相当于要分类的段的全长。使用Adam optimizer训练网络，学习率为0.001，批大小为16。完整的训练周期为100个迭代周期，系统参数调整阶段训练周期为15。

不同于第二章中的GMM模型，由于卷积神经网络模型对样本量大小相对不敏感，且卷积神经网络模型训练时间远远大于GMM模型，因此本章实验不再使用4折交叉验证法。替代方法使用由大约30％的原始训练数据组成的验证集来完成，选择使得训练和验证集不具有来自相同位置的片段，并且两组都具有每个类别的数据，类别比例保持相等。在每个迭代周期之后在验证集上评估模型性能，选择性能最佳的模型。

### 3.5.2 评价指标

本系统的评价指标之一依然基于准确率，在此不再赘述。此外，为了准确分析音频场景分类系统在各场景下的差异，以便后期优化系统，引入了混淆矩阵（Confusion Matrix）作为评价指标之一。混淆矩阵是判定分类精度的一种方法。在混淆矩阵中，其列坐标代表预测的值，行坐标代表实际的值。通过引入混淆矩阵则可以清晰的判断出本系统模型的分类是否存在混淆。

### 3.5.3 参数调整

根据上一节中讨论的参数调整方案，现进行初步试验以得出系统最佳参数。参考依据为分类准确度和单位迭代周期。系统均使用开发集数据进行训练与测试，以确保以尽可能小的代价体现系统性能，最后的整体实验阶段可能需要比参数调整阶段更长的单位迭代周期。

1. 滤波器数量的调整

滤波器数量的调整策略见表3-2：

#### 表3-2 滤波器数量调整策略

|  |  |  |
| --- | --- | --- |
|  | CNN\_1层 | CNN\_2层 |
| 方案a1 | 16 | 32 |
| 方案a2 | 32 | 64 |
| 方案a3 | 64 | 128 |
| 方案a4 | 128 | 256 |
| 方案a5 | 256 | 512 |

我们分别设计了4组方案以对照，滤波器数量呈倍数增加，且整个网络中其他参数固定。鉴于之后要调整滤波器尺寸，从参数较少的设计出发，滤波器尺寸暂定取3\*3。

#### 表3-3 不同滤波器数量的分类结果

|  |  |  |
| --- | --- | --- |
|  | 准确率 | 单位迭代周期 |
| 方案a1 | 45.7% | 0:00:15 |
| 方案a2 | 61.7% | 0:00:29 |
| 方案a3 | 72.8% | 0:01:00 |
| 方案a4 | 78.7% | 0:02:13 |
| 方案a5 | 77.9% | 0:05:23 |

根据上述实验结果可以发现，方案a4可以达到准确度与训练时间的最佳平衡。滤波器数量的无限制增加不仅没有提高准确率，且大大增加了单位迭代周期。因此，通过本轮实验决定最终训练模型CNN\_1层滤波器数量为128，CNN\_2层滤波器数量为256。

1. 滤波器尺寸的调整

滤波器尺寸的方案为4种，不同的滤波器尺寸决定了局部感受野的大小，也即确定了要提取特征的层次。通过调整参数，找到与模型最匹配的特征粒度，从而使分类准确率最大化。

#### 表3-4 滤波器尺寸调整策略

|  |  |  |
| --- | --- | --- |
|  | CNN\_1层 | CNN\_2层 |
| 方案a4 | 3\*3 | 3\*3 |
| 方案b1 | 5\*5 | 5\*5 |
| 方案b2 | 7\*7 | 7\*7 |
| 方案b3 | 9\*9 | 9\*9 |

第二轮参数调整的分类结果如表3-5所示：

#### 表3-5 不同滤波器尺寸的分类结果

|  |  |  |
| --- | --- | --- |
|  | 准确率 | 单位迭代周期 |
| 方案a4 | 78.7% | 0:02:13 |
| 方案b1 | 85.9% | 0:03:01 |
| 方案b2 | 89.1% | 0:04:15 |
| 方案b3 | 89.3% | 0:05:50 |

观察表3-5我们可以发现，当滤波器尺寸设置为7\*7与9\*9时，系统都能达到相对高的分类准确率，且方案b3分类性能略微优于b2。但考虑到两方案分类准确率差距甚微，且方案b3由于滤波器尺寸更大致使参数增加进而导致单位迭代周期有明显的增多，在最后的整体实验阶段会导致训练时间大幅增加。因此综合考虑分类准确率与单位迭代周期，本文决定采用效能比最佳的方案b2，即7\*7的滤波器尺寸。

在确定了滤波器的尺寸与数量这两个关键性因素后，最终的系统网络参数如表3-2所示：

#### 表3-6 卷积神经网络参数

|  |  |  |
| --- | --- | --- |
| **层（类型）** | **输出尺寸** | **参数（个）** |
| 卷积层\_1 | (40,500,128) | 6400 |
| 批标准化\_1 | (40,500,128) | 160 |
| 激活函数\_1 | (40,500,128) | 0 |
| 最大池化\_1 | (8,100, 128) | 0 |
| Dropout\_1 | (8,100, 128) | 0 |
| 卷积层\_2 | (8,100,256) | 1605888 |
| 批标准化\_2 | (8,100,256) | 32 |
| 激活函数\_2 | (8,100,256) | 0 |
| 最大池化\_2 | (2,1,256) | 0 |
| Dropout\_2 | (2,1,256) | 0 |
| 拍平 | 512 | 0 |
| Dense(全连接层)\_1 | 100 | 51300 |
| Dropout\_3 | 100 | 0 |
| Dense(全连接层)\_2 | 15 | 1515 |
| 参数总计：1,665,295 | | |
| 可训练参数：1,665,199 | | |
| 不可训练参数：96 | | |

### 3.5.4 实验结果与分析

根据3.5.3节的参数，在两个数据集上训练得到的结果分类准确率结果如表3-3所示：

#### 表3-7 音频场景分类准确率结果（单位：%）

|  |  |  |
| --- | --- | --- |
| **音频场景** | **开发集** | **验证集** |
| 沙滩 | 92.0 | 41.7 |
| 公交 | 94.6 | 26.9 |
| 咖啡/餐馆 | 89.1 | 55.6 |
| 汽车 | 99.0 | 70.4 |
| 市中心 | 100.0 | 91.7 |
| 森林小径 | 99.7 | 92.6 |
| 杂货铺 | 88.8 | 49.1 |
| 家 | 95.2 | 64.8 |
| 图书馆 | 89.1 | 37.0 |
| 地铁站 | 96.5 | 91.7 |
| 办公室 | 100.0 | 86.1 |
| 公园 | 91.3 | 21.3 |
| 小区 | 86.9 | 56.5 |
| 火车 | 87.8 | 60.2 |
| 电车 | 90.4 | 42.6 |
| 平均 | 93.4 | 59.2 |

分析表中我们可以发现，开发集的系统性能为93.4％，且各类的识别准确率差异不大，皆处在很高的水平上。验证集的平均性能为59.2％，说明系统泛化能力一般。而造成这种结果的原因很大一部分在于卷积层中的参数达百万之多，对训练的数据产生了记忆效应，因此在开发集上表现的很好，但在验证集上大打折扣。因此，后期对网络优化阶段应从改变网络结构入手，即不需要单纯的依靠参数的设置来提升系统的分类准确率。

通过比较各个场景类别的系统性能，我们注意到大多数场景相似的音频对的识别准确率非常相似，如公园与街道交通、地铁与电车。最难识别的是公共广场，在开发集中识别准确率最低，验证集​​中甚至更低（34.6%），而具有交通场景的街道与公园获得的最佳性能均超过80％。

混淆矩阵如表3-8所示：

#### 表3-8 音频场景分类结果的混淆矩阵

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 沙滩 | 公交 | 咖啡 | 汽车 | 市中 | 森林 | 杂货 | 家 | 图书 | 地铁 | 办公 | 公园 | 小区 | 火车 | 电车 |
| 沙滩 | 45 | 0 | 27 | 0 | 1 | 1 | 4 | 0 | 4 | 0 | 0 | 5 | 21 | 0 | 0 |
| 公交 | 2 | 29 | 17 | 19 | 6 | 0 | 4 | 2 | 0 | 5 | 0 | 0 | 1 | 9 | 14 |
| 咖啡 | 1 | 0 | 60 | 0 | 0 | 1 | 6 | 3 | 3 | 26 | 0 | 0 | 0 | 0 | 8 |
| 汽车 | 0 | 0 | 0 | 76 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 11 | 18 |
| 市中 | 0 | 0 | 0 | 0 | 99 | 3 | 1 | 0 | 0 | 0 | 0 | 3 | 2 | 0 | 0 |
| 森林 | 1 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 1 | 5 | 1 | 0 | 0 |
| 杂货 | 7 | 0 | 23 | 0 | 0 | 15 | 53 | 1 | 1 | 7 | 0 | 0 | 0 | 0 | 1 |
| 家 | 4 | 0 | 0 | 0 | 0 | 5 | 0 | 70 | 11 | 0 | 18 | 0 | 0 | 0 | 0 |
| 图书 | 4 | 0 | 0 | 0 | 0 | 5 | 3 | 8 | 40 | 13 | 35 | 0 | 0 | 0 | 0 |
| 地铁 | 0 | 0 | 0 | 0 | 6 | 3 | 0 | 0 | 0 | 99 | 0 | 0 | 0 | 0 | 0 |
| 办公 | 1 | 0 | 0 | 0 | 0 | 8 | 0 | 4 | 2 | 0 | 93 | 0 | 0 | 0 | 0 |
| 公园 | 3 | 0 | 0 | 0 | 6 | 20 | 4 | 0 | 0 | 0 | 0 | 23 | 52 | 0 | 0 |
| 小区 | 1 | 0 | 0 | 0 | 15 | 27 | 0 | 0 | 1 | 0 | 0 | 3 | 61 | 0 | 0 |
| 火车 | 3 | 5 | 0 | 10 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 65 | 19 |
| 电车 | 3 | 5 | 25 | 5 | 2 | 0 | 0 | 0 | 0 | 12 | 0 | 3 | 0 | 11 | 46 |

通过观察表3-8的混淆矩阵可以更深入地了解哪些类大多被错误分类。可以发现混淆度较高的分类情况在于公园-小区、办公室-图书馆、公交-电车、森林-小区等，这也与听觉常识相吻合。这可能表明本系统的模型更多地依赖于序列的背景噪声而不是音频事件的发生。此外，由于预处理过程中将双通道音频压缩成单通道，人为的减少了可分类因素，也可能是部分相似场景下混淆度较高的原因。

最后，第二章与第三章系统的对比如表3-9所示：

#### 表3-9 GMM系统与CNN系统的分类性能比较

从表中我们可以看出，在开发集上卷积神经网络模型的分类性能大幅优于GMM模型，但在验证集中两系统的分类性能各有优劣。

## 3.6 本章小结

本章首先介绍了基于卷积神经网络的音频场景分类系统的整体结构，然后从卷积神经网络的结构出发，介绍了其特点以及参数学习的方法；接着分别探讨了卷积神经网络各层应用到音频场景分类的可行性；再从系统设计的角度探讨了分类系统的实现细节。最后，先进行了参数调整找到了系统最优分类参数，然后进行了完整实验，验证了将卷积神经网络应用在音频场景分类中的可行性，也分析了其中的不足，并指出了可能的原因。

# 第4章 音频场景分类系统的改进

## 4.1 引言

在上一章中，我们通过实验验证了卷积神经网络在音频场景分类中应用的可行性，并在开发集中取得了大幅优于基线系统的效果。而之前在验证集中出现的泛化能力问题，本章将尝试从特征提取和网络结构优化两方面出发，以进一步提升音频场景分类系统在各种数据集上的分类准确率。

在第二章与第三章的系统中，特征提取部分一直采用的是MFCC特征。尽管MFCC可以用十几个系数简洁地描述特征，但是，能否改进MFCC使其能在诸如图书馆这样的大声场等其他复杂的环境下更好的胜任音频场景分类是本章要研究的一个方面。另外，上一章中的卷积神经网络结构在训练过程中也存在过于依赖训练参数的问题。目前得以广泛应用的性能良好的卷积神经网络框架有很多，能否将其结构稍加修改利用与音频场景分类之中以提高系统性能也是本章的一个着重点。

基于以上的讨论，本章将先从音频处理、特征提取方面出发，再研究改进的卷积神经网络结构设计。最后，将改进后的系统应用于实验中，以检测最终的分类准确率，并与前两章进行对比。

## 4.2 音频处理

本节介绍了所提出的音频预处理方法。还描述了所提出的ConvNet架构的部分细节以及如何从它们配置集合模型。本章还使用了多个版本的梅尔频谱图，它们可以大致分为三种方法，即双耳表示法，谐波冲击源分离法和背景减法。下面给出了每种方法的详细解释，以及提取梅尔的例子。

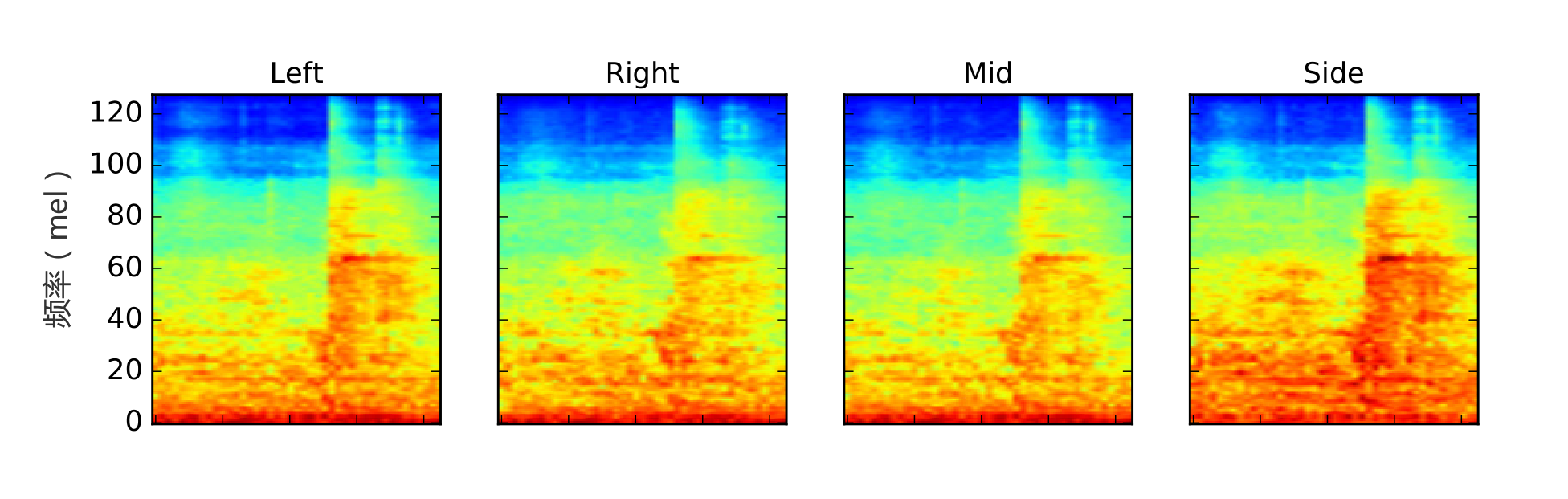
### 4.2.1 双耳表示法

尽管在录音时使用立体声设备录制很常见，但通常在音频处理时都会在处理之前对信号求平均来使其成为单声道。固然使用单声道便于处理且易于特征提取，然而这样会丢失不少空间信息。如果重要的音频信息仅在其中一个通道中捕获良好，则可能会出现问题。因为将双通道平均到一个通道后会降低信噪比，这样两个声道的差别不容易被体现出来，因此在分类过程中极易造成混淆。分别对两个通道进行分析可以缓解这个问题。鉴于使用双耳表示法在先前DCASE挑战中取得的优异结果（Eghbal-Zadeh等，2016），本文决定在特征处理过程中使用LR表示法（Left、Right，即左右）和MS表示法（Mid、Side，即中间—侧面）。

LR表示法代表的是常规立体声录音中的左声道与右声道。例如，汽车在麦克风前经过，声音从L声道移动到R声道或从R声道移动到L声道，这种在单声道中只体现为幅度变化。通过引入LR表示法，可以体现出音源在空间中的移动。应用在本章中仅将源音频文件的左右声道分离即可。

而MS表示法则强调到达立体声麦克风的每一侧的声音之间的时间差。其中，Mid通道定义为L + R，Side通道定义为L-R，即两个通道之间的差异。

分别将对以上四种表示法提取MFCC特征，其谱图如图4-1所示：



#### 图4-1 双耳表示法的MFCC频谱图

对于LR谱图和MS谱图，分别将其输入4.3节的卷积神经网络中，作为分类特征之一。

### 4.2.2 谐波冲击源分离法

声音一般可分为两种类型：谐波和冲击声。如[18]中所述，在传统的研究工作中，谐波冲击源分离（HPSS）算法是在音乐信号处理的背景下提出的，目标是将输入音频信号分解为由所有谐波和由所有冲击声组成组成的信号。

在这里，我们使用基于NMF的HPSS算法[19]将数据集中的音频剪辑分成两部分，这使得能够分别利用声音的谐波和打击乐方面。在分离之前，立体声声音被转换为单声道。

用于分离的实验参数分别为a，B，y和0.7,1.05,1.05和0.95，并且帧大小和跳跃大小分别为4,096和1,024个样本。基数的总数设定为200，由100个平坦初始化的冲击基础和100个随机初始化的谐波基础组成。维纳过滤不用于NMF的后处理，但是，我们已经完成了100次迭代中的最后30次迭代，不包括先前的拼版以减少在分离过程中可能产生的任何伪像。

### 4.2.3 背景减法

通常，中值滤波用于去除扫描图像中的噪声。 Moore Jr.和Jorgenson [20]使用这种技术通过从原始数据中减去中值滤波数据来进行对象提取。

虽然这种技术在图像处理领域更常用，但我们认为消除“稳定”是有用的。

来自环境或录音设备的噪音。通过这样做，我们期望梅谱图中的声事件的光谱特性被强调并且对于过度拟合更加鲁棒。与物体检测技术类似，我们对梅谱图应用中值滤波，并从原始版本中减去。我们首先在处理之前将立体声音频转换为单声道。用于中值滤波的滤波器大小对于时间轴是21,43,87（大约0.5s，1.0s和2.0s），并且对于频率轴是1,11，其通过实验凭经验选择。

请注意，在频率轴上使用1的内核大小实际上是随时间推移的1-D中值滤波。如图1的底行所示，背景减法过程强调来自相邻区域的不同光谱特征，这使得更容易检测声事件。

## 4.3 改进系统的结构设计

### 4.3.1 网络结构

### 4.3.2 网络组织方式

## 4.4 实验准备与结果

## 4.5 本章小结

# 第5章 结 论

# 参考文献

Eronen A J, Peltonen V T, Tuomi J T, et al. Audio-based context recognition[J]. IEEE Transactions on Audio Speech & Language Processing, 2006, 14(1):321-329.

Malkin R G, Waibel A. Classifying user environment for mobile applications using linear autoencoding of ambient audio[C]// ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing. 2005.

Krijnders J, t Holt G. Tone-fit and MFCC scene classification compared to human recognition[J]. Energy [dB], 2013, 400(450): 500.

Sawhney N, Maes P. Situational awareness from environmental sounds[J]. Tech-nical Report, Massachusetts Institute of Technology, 1997.

Clarkson B, Sawhney N, Pentland A. Auditory context awareness via wearable computing[J]. Energy, 1998, 400(600): 20.

Patil K, Elhilali M. Multiresolution auditory representations for scene classification[J]. cortex, 2002, 87(1): 516-527.

Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

Ballas J A. 1993. Common factors in the identification of an assortment of brief everyday sounds[J]. Journal of experimental psychology: human perception and performance, 19(2): 250.

Peltonen V T K, Eronen A J, Parviainen M P, et al. 2001. Recognition of everyday auditory scenes: potentials, latencies and cues[J]. PREPRINTS-AUDIO ENGINEERING SOCIETY.

Dubois D, Guastavino C, Raimbault M. 2006. A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories[J]. Acta acustica united with acustica, 92(6): 865-874.

Tardieu J, Susini P, Poisson F, et al. 2008. Perceptual study of soundscapes in train stations[J]. Applied Acoustics, 69(12): 1224-1239.

Eronen A, Tuomi J, Klapuri A, et al. 2003. Audio-based context awareness-acoustic modeling and perceptual evaluation[C]//Acoustics, Speech, and Signal Processing, 2003.

Lin M, Chen Q, Yan S. 2013. Network in network[J]. arXiv preprint arXiv:1312.4400.

Davis S，Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences [J]. IEEe transactions on acoustics, speech，and signal processing, 1980, 28 (4): 357-366.

Lidy T, Schindler A. 2016. CQT-based convolutional neural networks for audio scene classification[C]//Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016). DCASE2016 Challenge, 90: 1032-1048.

Aytar Y, Vondrick C, Torralba A. 2016. Soundnet: Learning sound representations from unlabeled video[C]//Advances in Neural Information Processing Systems. 892-900.

Santoso A, Wang C Y, Wang J C. Acoustic scene classification using network-in-network based convolutional neural network[R]. DCASE2016 Challenge, Tech. Rep, 2016.

Reynolds D A, Rose R C. Robust text-independent speaker identification using Gaussian mixture speaker models[J]. IEEE transactions on speech and audio processing, 1995, 3(1): 72-83.

Dempster A . Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society, Series B, 1977, 39.

Nuttall A H . Some Integrals Involving the Q-Function[J]. IEEE Transactions on Information Theory, 1972, 21(1):95-96.

Rijsbergen C J V. Information Retrieval[M]. 1979.

Wiesel T N , Hubel D H . EXTENT OF RECOVERY FROM THE EFFECTS OF VISUAL DEPRIVATION IN KITTENS[J]. Journal of Neurophysiology, 1965, 28(6):1060-1072.

Fukushima K. Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position[J]. Biological Cybernetics, 1980, 36(4): 193-202.

Hinton G E, Osindero S, Teh Y W, et al. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554.

Salakhutdinov R, Mnih A, Hinton G E, et al. Restricted Boltzmann machines for collaborative filtering[C]. international conference on machine learning, 2007: 791-798.

Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. arXiv preprint arXiv:1502.03167, 2015.

Shimodaira H. Improving predictive inference under covariate shift by weighting the log-likelihood function[J]. Journal of statistical planning and inference, 2000, 90(2): 227-244.

|  |
| --- |
|  |

Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. arXiv preprint arXiv:1207.0580, 2012.

Eghbal-Zadeh H, Lehner B, Dorfer M, et al. CP-JKU submissions for DCASE-2016: a hybrid approach using binaural i-vectors and deep convolutional neural networks[J]. IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE), 2016.

# 致 谢