硕士学位论文

基于深层神经网络的音频特征提取及场景识别研究

RESEARCH ON AUDIO FEATURE EXTRACTION AND CONTEXT RECOGNITION BASED ON DEEP NEURAL NETWORKS

王乃峰

哈尔滨工业大学

2015 年 6 月

国内图书分类号：TP391.4　　学校代码：10213
国际图书分类号：681.3　　密级：公开

工程硕士学位论文

基于深层神经网络的音频特征提取及场景识别研究

硕 士 研 究 生： 王乃峰

导　　师： 郑铁然副教授

申 请 学 位： 工程硕士

学　　科： 计算机技术

所 在 单 位： 计算机科学与技术学院

答 辩 日 期： 2015 年 6 月

授予学位单位： 哈尔滨工业大学

Classified Index: TP391.4
U.D.C: 681.3

Dissertation for the Master Degree in Engineering

RESEARCH ON AUDIO FEATURE EXTRACTION AND CONTEXT RECOGNITION BASED ON DEEP NEURAL NETWORKS

Candidate: Wang Naifeng
Supervisor: Associate Prof. Zheng Tieran
Academic Degree Applied for: Mater of Engineering
Speciality: Computer Technology
Affiliation: School of Computer Science and Technology
Date of Defence: July, 2015
Degree-Conferring-Institution： Harbin Institute of Technology

摘 要

音频场景识别(Audio Context Recognition)是人工智能领域重要的研究方向之一，该技术依据周边声音感知环境动态，对机器作出进一步智能选择有着非常重要的意义。近年来有较多的学者涉足这一研究方向，他们大多采用先特征提取后分类器分类的研究框架，其中，对如何提取能够反映音频场景声学特性的识别特征方面给与了较多的关注。所采用的声学特征可以粗略的分为两大类：短时特征和长时特征。短时特征如单音轨梅尔频率倒谱系数、多音轨梅尔频率倒谱系数、梅尔频率倒谱系数和稀疏特征的联合特征等；长时特征多为音频段的长时统计值或基于语义相关性的特征等。从研究结果上看，目前的声学特征都有其不足之处，短时特征往往不足以完整地刻画一个音频场景的特性，长时特征往往缺乏对音频段内部细节的描述，而长时统计值中缺失的内部结构信息对区分音频场景也有重用的价值。本文对既能反映音频段长时特性又能反映局部结构性的声学特征的提取方法进行研究，并验证了它们在音频场景识别任务中的有效性。

深层神经网络能够通过自学习来发现适合分类任务的特征向量，这已经在图像尤其是自然图像的结构性特征分析方面得到了验证，这些特征提取方法能够很好地反映出图像的结构信息，相对于主观分析方法有着很大的优势。所以，本课题借助深层神经网络的特征分析能力在语谱图上进行场景长时结构性特征的分析与提取。主要研究内容如下：

首先研究了基于卷积神经网络的音频场景特征分析提取方法，卷积神经网络通过卷积和下采样操作对输入数据进行非线性映射,最终通过重构误差的反向传播进行参数的调节，从而提取出能够刻画音频场景特性的声学特征。卷积神经网络在训练是以输入数据的类别重构误差最小化为优化目标的，所以训练过程需要大量的有标签数据。

现实情况下，往往难以得到大量带标签的训练数据，因为对数据进行标注费时费力，所以本文也提出了基于解卷积神经网络的特征提取方法。解卷积神经网络模型在特征分析处理过程中不仅保留了卷积神经网络的卷积操作和下采样操作，而且还在原来的基础上有了些改进，其参数更新过程是基于对输入数据重构误差的反向传播进行的，这样就不需要带标签的数据。采用这种特征提取方法，音频场景数据的采集相对来说就较为容易，训练数据不足的问题就得到了很好的解决。

通过实验验证，我们得到的结果是，采用基于卷积神经网络得到的音频特

征，进行场景识别的性能有很大提升；基于解卷积神经网络分析得到的特征虽然对识别性能没有较大的提升，但是针对训练数据不足的问题它还是有效的。

关键词：音频特征提取；卷积神经网络；解卷积神经网络；音频场景识别

Abstract

As one of the artificial intelligence research orientation, the audio context recognition can apperceive the environment dynamic information according to nearby sounds, that is very important for a further intelligent choice of which the machine to make. Recent years, there are many researchers have focused on this area, yet most of them did their research by the following framework: what the first step is the feature extraction, and nextly is the classification of the pattern, among of them have paid much attention to how to extract the recognition feature which can reflect the acoustic properties of the audio context. The acoustic features can be roughly divided into two main classes: short time feature and long time feature. Short time feature includes the following content: single track mel frequency cepstral coefficients、 multiple tracks mel frequency cepstral coefficients、 mel frequency cepstral coefficients and the combination feature of the sparse feature. Most of the short time feature are statistic value of long time audio segment or feature based on the semantic correlation. From the result, we can see that all the above features have its' shortage, short time feature can not fully describes the acoustic context, and yet long time feature may lack of the description of the inside detail information of the acoustic segment, which is very important for the classification of the audio context. This paper focus on the feature extraction method which aims to find the ones that can reflect not only the long time properties but also the local structural acoustic properties, and proves their efficience for the acoustic context recognition by our experiences.

We can get the best suitable features for classification by the deep neural network which can learn by itself. It has been proved in image structural feature analysis, especially the nature image, it can reflect the structural information of images better than people's subjective analysis. So, we will do the long time feature analysis in audio context spectrogram with the help of feature analysis ability of deep neural network. Contents of this paper is arranged like the following:

Firstly we do the research about audio context recognition method based on convolutional neural network, namely CNN. We can do nonlinear mapping on the input data by convolution and down sampling operation of CNN, and we can update the parameters by back propagation of reconstruct errors, so that we can extract the audio features which can describe the audio context properties. The feature analysis and the classification are seemed as two parts of the whole training process, which is supervised so that we have to supply lots of labeled data.

In fact, the labeled data is so difficult to get which may cost lots of time and

manual labor that we also introduce another feature analysis method， a method based on the deconvolutional neural networks, namely DeCNN. Not only the convolution but also the down sampling operation are stayed in DeCNN, what's more there has some progress than CNN, what content is that the networks parameters update procession is based on the back propagation with respect to the reconstruction error of input data rather than the labels corresponding to the input data. So that, we will do not need so many labeled data. By this way, the unlabeled data is so easy to get, the lack problem of the input labeled data can be solved.

Experience result shows that the long time structural feature based on CNN has big progress in the audio context recognition rate than the baseline system. Though the result of DeCNN is not so good, yet DeCNN solves the lack problem of labeled training data.

Keywords: Audio feature extraction, CNN, DeCNN, Audio context recognition

目 录

摘 要.....I
ABSTRACT.....III
第 1 章 绪 论.....1
1.1 课题背景及研究的目的和意义.....1
1.2 国内外研究现状.....1
1.2.1 声学特征分析研究现状.....1
1.2.2 音频场景识别研究现状.....3
1.2.3 深度神经网络的研究现状.....4
1.3 研究内容.....6
第 2 章 基于 MFCC 和 KNN 的场景识别基线系统.....8
2.1 引言.....8
2.2 基本流程、预处理及特征提取.....8
2.2.1 基线系统基本流程.....8
2.2.2 音频信号预处理及特征提取.....8
2.3 场景识别.....10
2.3.1 K 近邻算法的理论基础及算法流程.....11
2.3.2 KNN 模型及基本要素.....12
2.3.3 K 近邻算法应用.....15
2.4 实验结果及分析.....16
2.4.1 实验数据.....16
2.4.2 实验参数调整及结果分析.....17
2.5 本章小结.....17
第 3 章 基于卷积神经网络的音频特征提取及场景识别.....19
3.1 引言.....19
3.2 现有音频特征在场景识别方面的适用性分析.....19
3.3 基于 CNN 的深层特征提取原理及学习算法.....20
3.3.1 CNN 拓扑结构.....20
3.3.2 CNN 计算方式.....21

3.3.3 CNN 学习算法.....22
3.4 基于 CNN 的场景特征提取及识别分析.....25
3.4.1 语谱图的特性以及 CNN 方法的适用性.....25
3.4.2 卷积滤波器的设计.....27
3.4.3 特征提取及分类.....28
3.5 实验过程及结果分析.....28
3.5.1 实验数据.....28
3.5.2 实验基本网络结构.....29
3.5.3 实验参数调整及结果分析.....29
3.5.4 实验流程及中间结果分析.....33
3.6 本章小结.....37
第 4 章 基于解卷积神经网络的音频特征提取及场景识别.....38
4.1 引言.....38
4.2 基于解卷积神经网络的特征分析.....38
4.2.1 解卷积神经网络的拓扑结构.....39
4.2.2 解卷积神经网络的计算方式.....39
4.2.3 解卷积神经网络的学习算法.....43
4.3 解卷积神经网络用于音频场景的特征分析及识别算法.....44
4.3.1 解卷积神经网络的适用性分析.....44
4.3.2 解卷积神经网络用于音频场景特征提取及识别算法.....46
4.4 实验结果及分析.....46
4.4.1 实验数据预处理.....46
4.4.2 实验基本网络结构.....46
4.4.3 实验参数调整及结果分析.....47
4.5 本章小结.....48
结 论.....50
参考文献.....51
哈尔滨工业大学学位论文原创性声明和使用权限.....55
致 谢.....56

第 1 章 绪　论

1.1 课题背景及研究的目的和意义

声音感知和理解作为人工智能领域的一个重要的研究方向，在过去的时间里受到了很多专家和学者的关注。特别是最近一段时间，随着计算机科学的进步与发展，其研究工作更是备受青睐。在此研究领域中，对语音的识别和理解任务已经得到了广泛关注,而非语音环境声音的感知和理解研究也有重要的意义。

音频场景识别项技术是通过对周边环境声音的理解来对环境的动态改变进行感知识别，它更多的涉及到对非语音信号的感知和理解，其对机器做出进一步的智能选择有着非常重要的指示作用，此外它还可以用在军事、刑侦等领域。音频场景识别性能的好坏很大程度上依赖于识别过程中所使用的声学特征。

现在已有的场景识别方法所用到的音频特征大都是基于倒谱域的 MFCC，此外还有其它频域、时域方面的特征。但是它们或者是短时特征或者是长时特征的统计值，短时特征不能完整地对音频场景进行描述，而长时统计特性会丢失场景信号的局部结构性信息，最终都会影响音频场景的识别效果。所以，长时结构性特征对音频场景识别尤其重要。

1.2 国内外研究现状

在语音信号处理领域音频特征分析、提取的研究工作从未中断过，此外越来越多的学者开始关注音频场景识别方面的探索研究，同时也取得了一定的成绩。本课题尝试利用深度神经网络在语谱图上进行音频特征分析、提取，进而进行音频场景的分类。下面从声学特征分析、场景识别和深层神经网络等方面详细阐述下国内外的研究现状。

1.2.1 声学特征分析研究现状

按照被分析音频所采用的表示形式的不同，可以将音频特征提取分成：时域信号分析、频域信号分析、倒谱域信号分析。

（1）时域声学特征

音频信号无论是模拟信号还是经过采样、量化后得到的数字信号，其原始的表示形式都是以时间为自变量的时域信号。时域信号的波形最直接最简单，

分析起来简单直观、计算量小，而且对应的物理意义明确，所以时域分析法是音频特征分析研究历程中应用最早且范围较广的方法之一。时域分析法中的短时能量分析是从能量角度进行的，短时自相关系数和短时平均幅度差分析都是从时域相似度角度进行的,短时过零率则在某种程度上反映了信号的频率特性。

短时能量分析是基于语音信号随着时间变化其能量值的变化较剧烈这一性质进行的。短时能量的形态很大程度上取决于所选择的窗函数的形式，而且短时能量值等于音频信号的幅度值的平方，这样也就人为地增加了高低信号之间的差距，这在有些场合是不适用的。

短时平均过零率就是在较短时间内音频信号对应的时域轨迹线穿过零值那条直线的次数。很明显，通过短时平均过零率我们可以得到信号的一些频率方面的信息，同时我们也看到低频对短时平均过零率的影响还是挺大的。故后来就在此基础上进行拓展，设定了一个下限阈值，即将通过零值改为通过该阈值，经过修改之后的方法具有一定的抗干扰能力。

短时自相关函数在短时处理的基础上可以反映信号在时域内的相似程度。通关自相关函数我们可以得到信号自身的一些性质，比如说同步性和周期性。但是短时自相关函数的计算过程中会产生很大的计算量，因为过程中涉及到大量的乘法运算。为此，人们又提出了短时平均幅度差函数，它是对周期信号做差分运算，这样不仅可以用于基音周期的检测，而且计算更简单。

（2）频域声学特征

时域分析法虽然简单易懂，但是时域信号含有的有效信息不足。相对而言，频谱分析和人类听觉系统有更加密切的关系，所以频谱含有更多的声学特性和感知特性，而且频域分析法对外界环境变化更加鲁棒，所以音频特征分析更多的是围绕着频域进行的。

研究者广泛采用的频域声学特征有：滤波器组分析方法，傅里叶变换法、线性预测分析方法等

滤波器组分析方法是较早的频域分析的方法之一，它是通过一组带通滤波器，输出具有一定中心频率的子带信号。该方法不仅简单实用而且对外界环境变化更加鲁棒。

傅里叶变换法是非常经典的工具分析方法，不仅可以用来作为语音信号的处理方法。傅里叶变化就是简单的正交变换，变换的基底是复指数形式的函数，这种变化在理论和计算两个层面都具有很大的优越性。又因为音频信号具有随着时间变化自身变化较缓慢这一特性，所以频域分析和时域分析一样也是进行短时分析的，短时傅里叶分析方法相当于带通滤波器的作用。

线性预测分析(Linear Predication Coding, LPC)方法在语音信号处理领域

应用范围很广。它的主要思想是通过之前的几个音频信号的采样值来预测当前信号的采样值，预测结果和真实值之间肯定有差距。以最小均方误差为优化目标，我们会得到一组加权组合系数，这就是线性预测分析系数。

（3）倒谱域特征

倒谱分析是在对语音信号进行解卷积处理过程中用到的一种非参数方法，倒谱域特征可以通过对频谱数据取对数然后再进行反傅里叶变换得到。倒谱域特征有：LPCC、MFCC 和 PLPCC 等。

LPCC 是在线性预测分析的基础上取对数，然后再做反傅里叶变换得到的一种音频倒谱特征系数,我们可以把它看做是对原始信号短时倒谱的一种近似。

目前应用最为广泛的音频特征是 MFCC，它与其它音频特征的不同之处在于有效的结合了人耳的听觉感知特性和语音的产生机制。人的耳蜗具有类似非线性滤波器一样的滤波作用，即它是在对数频率尺度上进行滤波的。这启发人们尝试把常规的频率进行非线性处理，向对数频率上面映射，最终使语音识别等多个处理工作的性能得到了很大的提升。

PLP 是通过使用简单的模型来模拟人耳的多种听觉特性，得到更加符合人耳听觉特性的音频特征，但是其计算量较大，PLPCC 则是在此基础上取对数然后进行反傅里叶变换得到的感知线性预测倒谱系数。

（4）音频场景识别应用中的声学特征

现有的音频场景识别方法所使用的声学特征多数为 MFCC。除了上述声学特征，频域特征还有子带能量比、基频等；倒谱域特征还有对数频率倒谱系数等；Selina chu[1]提出基于时频分析的音频特征进行场景识别；此外，Agostini[2]等人还提出使用长时特征的统计值进行音频场景的识别。但是这些特征或者是短时特征或者是长时统计特征,短时特征不能完整地刻画音频场景的声学特性，而长时特征在进行统计值计算时会丢失局部性信息。

1.2.2 音频场景识别研究现状

音频场景可以看做是由一个或几个特定类别的声学事件构成的音频段。它往往通过对特定声学事件的检测来对整个音频场景所属类别做出判断。例如浪花拍打沙滩的声音和海面上各种海鸟的叫声等一起构成了海滩这个特定的音频场景，要识别这一特定场景首先要检测出其中的特定声学事件如海鸟的叫声，然后才能对其场景类别做出判断。

到目前为止，音频场景识别领域已有识别的方法都是按照模式识别框架进行的。即首先进行音频场景数据的特征分析、提取，然后进行场景的分类。常用到的识别器有支持向量机、高斯混合模型、k 近邻模型和隐马尔科夫模型等。

此外还有人尝试使用模型组合的方式进行音频场景的分类，Wei-Ta chu[3]等人通过组合支持向量机和高斯混合模型来进行音频场景的分类，Khin Myo chit[4]把隐马尔科夫模型和支持向量机进行级联来进行场景识别，最终都取得了较好的效果。

Toni Heittola[5]等人提出基于关键声学事件直方图的场景识别方法。大致思路如下：首先进行关键声学事件的检测，然后根据检测出的声学事件对场景进行建模，最后根据得到的场景模板进行识别。Virtanen[6]提出了使用声源分离的方式进行关键声学事件的检测，大致思路为：利用非负矩阵分解把声源信号进行分解，得到几个独立的音轨信号；然后在每个独立的信号上面按照上面的处理方式进行事件的检测，最终进行音频场景的分类，取得了较好的识别效果。由于声学事件之间存在重叠且其先验知识相对匮乏，这种情况下的声学事件检测相对就会困难很多，为此 Mesaros[7]等人提出了基于潜在语义分析的声学事件检测的方法来进行重叠事件的相关性分析，最终很好地解决了由于声学事件之间存在重叠而造成场景识别性能低的问题。

Chu[1]等人提出使用匹配追踪的方式直接对音频场景信号进行分析，得到对于背景噪声来说更加鲁棒的特征。匹配追踪的目标是找寻一组最小数目的基底来对信号进行稀疏且有效的分解。经过匹配追踪分析，原始信号就可以用一组基底向量进行表示，即得到了新的特征向量，然后使用这种新的特征参数直接对音频场景进行分类。Peltonen[8]等人通过使用 MFCC 和子带能量比作为音频场景的声学特征，以高斯混合模型和 KNN 分类器作为场景的分类器进行最终的分类，此外 Guo[9]在前人研究的基础上，通过使用频域特征和支持向量机进行音频场景的识别。Ajmera[10]则是把基于音频场景类后验概率得到的动态特征用作场景声学特征，使用隐马尔科夫模型进行最终的场景识别。

Kyuwoong[11]等人使用 MFCC 作为场景识别的声学特征，然后通过使用高斯混合模型构建语义模型，得到语义模型直方图，最后使用 KNN 分类器进行最后的场景识别。SriniVasan[12]等人借助音频场景识别的结果对语音信号进行分离来提高最终的语音识别性能；Akinori[13]等人通过使用多阶 GMM 来对声学事件进行检测，然后进行音频场景的识别。

1.2.3 深度神经网络的研究现状

深度神经网络较传统的神经网络相比，其层次数目上有较多增加。神经网络的研究基础是感知机模型，它是一个简单的两层网络，类似于其它的二类分类器它只具有简单的线性分类的功能。因为人工智能的目标就是生产具有人类或者超越人类智商的机器来开展更高级的工作，所以一直以来人们对神经网络

的研究也是模仿人类的神经系统对事物的识别做出的反映，即感知事物、分析事物、最终做出判断这一思路。但是，人类的神经系统是非常复杂的，只是简单的视觉神经系统而言其处理工作都是分级进行的。从视网膜开始提取出低级特征如事物的边缘特征，经过复杂的处理组合到较高级别的特征即事物的基本形状或者目标事物的局部状态，最后是更高级的特征即整个目标事物。较高层次的特征都是低层粒度较小的特征经过非线性变换得到的，而且高层特征更加抽象更加概念化，所以其表达能力也相应的更加强。人类的这种神经系统处理机制为多层神经网络这个概念的提出提供了很好的启发性信息，即人类大脑的认知过程也是经过多层抽象多次非线性组合进行的。

感知机模型早在上个世纪就已经被提出[14,15]，虽然在很早之前也有人提出过多层神经网络这一概念，但是由于其理论分析难度较大，训练方法不仅需要很多经验而且训练过程计算量非常巨大，导致其研究工作在相当长的时间内没有进展。相反层次较浅的模型如支持向量机(support vector machine, SVM)[16-19]和增强模型(boosting)[19,20]以它们清晰的理论和简单的计算占得先机。但是，层次数的优势也成为了限制它们的根源。由于层次数目较少，在训练样本和计算单元受限的情况下其对复杂抽象的函数的表达能力相对来说也就相当有限。这类浅层模型最大的缺点是它们在进行分类识别的过程中都是使用主观得到的特征，这种靠人类经验和运气成分的工作不仅仅非常费力气而且其效果不会太好。这些问题在深层神经网络中都可以得到解决。

第一个被训练出的深层神经网络模型是 1989 年由纽约州立大学 Yahn LeCun[21]教授提出的卷积神经网络模型 lenet-5。但是由于种种原因深度神经网络的进一步发展一直不尽如人意，随着计算机科学与技术的不断发展尤其是并行处理技术的进步，当年因为计算量的缘故而被放置下的深度神经网络模型又出现在人们的视野当中。第一个引发深度学习研究热潮的是多伦多大学的 Geoffrey Hinton[22-24]教授，他在 2006 年提出的深度置信网络模型(Deep belief net, DBN)及相应的快速学习算法不仅很好地解决了之前由于随着层次数目的增加而产生的梯度消失问题，而且应用效果也取得了突破性进展[25]。它的训练方式与以往的深层网络不同的地方是整个模型的训练是逐层进行的，训练过程如下：首先通过预训练得到各层的参数初始值，然后在此基础上通过反向传播进行参数调优得到最终收敛的模型。后来 Ruslan Salakhutdinov[26-28]提出的深度波尔茨曼机(Deep Boltzmann Machine, DBM)进一步提升了深度神经网络的研究热度。2011 年微软语音组[29-31]通过用深度神经网络模型代替声学模型中的高斯混合模型来求隐马尔科夫链中状态的发射概率进而得到基于深度学习的声学模型，最终使整体的识别错误率降低了超过 20%，这一突破性成果是该领域近十几年

来最大的进步。2012 年 Krizhevsky[32]等人利用他们提出的深度卷积神经网络模型在 ImageNet 数据集上更是取得了历史最好识别成绩。现阶段由于数据量的井喷式增长，即使不进行预训练处理，通过增加网络的层次数深度神经网络也取得了相当好的性能。2013 年微软在语音识别领域又有新的进展，宣称研发出一种新型的语音识别技术，能够提供接近即时的语音向文本转化的服务，较目前的语音识别技术其速度提升了两倍之多，同时准确率也相应的提升了 15 个百分点。这项技术不仅是学术界的成果，而且还可以应用到工业界，像微软的搜索引擎 Bing 可以借助这项技术来提升通讯的速度和准确性。目前，这项技术已经在 Windows Phone 上面进行了测试。国内工业界搜索巨头百度公司也发力在深度学习上，不惜投入重金在硅谷组建了致力于深度学习研究的实验室。近年来有学者[33,34]开始尝试使用递归神经网络进行语音识别，2014 年百度首席科学家吴恩达领衔的团队宣称在收集到的几千小时的训练数据上通过加噪声处理得到了近十万小时的语料库，在此基础上进行深层递归神经网络的训练，得到了非常好的识别性能[35]。此外，国内的科大讯飞、中科院等公司或单位都相继进行了深度神经网络在语音识别领域的研究，特别是科大讯飞取得了世界领先的识别成绩。2014 年 Facebook 公司的深度学习项目 DeepFace[36]通过在当前最大的人脸库上训练深层神经网络进行人脸识别，最终取得了 0.9725 的识别率，这基本上可以和人类相比肩。此外，堆叠多层条件随机场等其它类型的多层神经网络模型也被用作语言类型识别、序列标注等任务中[37]。此外，通过卷积深度置信网络[38,39]进行音频特征分析，然后用在说话人识别、音素识别等任务中也取得了相当好的效果。

上面介绍到的所有深度神经网络的研究成果都是基于深度神经网络独特的特征分析抽取能力取得的。

1.3 研究内容

音频场景识别作为人工智能的一个研究方向，虽然受到了较多学者的关注，但是从音频场景识别这一研究领域取得的成绩看，其研究工作还远没有那么成熟。针对这一现状，本课题从场景识别的几个不同角度提出自己的想法。首先是特征提取方面，然后是音频场景识别方面。大体思路为：首先把特征分析角度从时域或者频域特征分析转变为基于频域的长时分析；深度神经网络在图像分析领域有着非常好的识别性能，所以我们借助其特征分析的能力对语谱图进行深层特征分析提取；然后我们使用分类器进行最后的分类，具体研究内容章节安排大致如下：

第二章介绍了本课题用作结果对比的基线系统，即基于梅尔频率倒谱系数

和 k 近邻分类器的实验系统。在这一章，我们首先介绍了基线系统的基本流程，然后对流程中的主要步骤进行展开，依次详细介绍了数据预处理工作、特征提取工作，最后介绍了根据提取到的特征向量进行音频场景分类的工作。实验过程中我们根据通过调节影响实验结果的参数得到的对应结果选出最佳的参数值。

第三章介绍了基于卷积神经网络的音频特征提取及场景识别算法的研究。在这一章里，我们首先分析了现有音频特征的不足之处和基于频域的长时分析的好处；然后进行卷积神经网络的适用性分析，以及卷积神经网络在语谱图上进行深层特征分析提取的有效性分析；紧接着介绍了卷积神经网络的计算方式和学习算法；最后进行实验系统的算法设计以及基于此算法进行了实验验证，并把实验系统结果和基线系统结果进行对比分析。

第四章介绍了基于解卷积神经网络的音频特征分析、提取以及最后的音频场景分类。本章我们首先介绍了解卷积神经网络的计算方式和学习算法，然后分析了解卷积神经网络对场景特征分析的适用性，最后根据解卷积神经网络的结构特性设计了本实验系统的算法。按照我们设计的算法进行实验，把最终得到的实验结果和基线系统的结果进行对比分析。

最后在结论中，我们对整篇论文进行了总结。首先分析了课题研究方向目前已经取得的成绩，以及存在的问题和能够提升的空间；然后给出了本文的章节总结，最后给出了本课题所研究的内容在理论层面的展望和实验过程中存在的不足和后续可以完善的地方。

第 2 章 基于 MFCC 和 KNN 的场景识别基线系统

2.1 引言

当前音频场景识别领域识别效果最好，应用最为广泛的音频特征是梅尔刻度倒谱系数(MFCC),这是由于它模拟了人类的听觉系统，对信号在原始频率(hz)上进行了非线性弯曲。K 近邻(k-nearest neighbor, KNN)分类器是目前应用较为广泛的多类别分类器，其原理简单、效果好，特别是随着训练样本数据的增多，效果会更好。所以，我们采用 MFCC 和 KNN 作为我们场景识别的基线系统。本章节我们首先给出基线系统的基本流程以及各个模块的详细展开，然后对实验结果进行分析得出结论，作为后面两个章节的对比基准。

2.2 基本流程、预处理及特征提取

2.2.1 基线系统基本流程

本课题的基线系统是基于 MFCC 和 KNN 构建的，大致流程如下：首先是音频信号的预处理工作，然后对预处理得到的中间结果进行特征分析，最后在得到的特征向量上进行音频场景分类。

音频信号预处理工作主要是对由音频设备采集到的原始模拟信号进行处理，常规操作有：预滤波、模数转换、预加重、分帧、加窗等；特征提取的目的是提取出相对来说最能体现出当前音频信号的内容数据，我们采用的特征是 MFCC,所以涉及到的工作有：快速傅里叶变换(Fast Fourier Transformation, FFT)、三角滤波、离散余弦变换(Discrete Cosine Transformation, DCT)、倒谱均值减、差分计算[40]等；最后一步是对特征数据进行分类，我们选择的 KNN 分类器无须进行训练，可以直接对特征向量进行分类操作。本课题的基线系统基本流程如图 2-1 所示。

2.2.2 音频信号预处理及特征提取

预处理阶段，主要有分帧、加窗操作。

2.2.2.1 分帧、加窗

声卡设备采集到的信号是模拟信号，经过数字化处理会得到对应的数字信

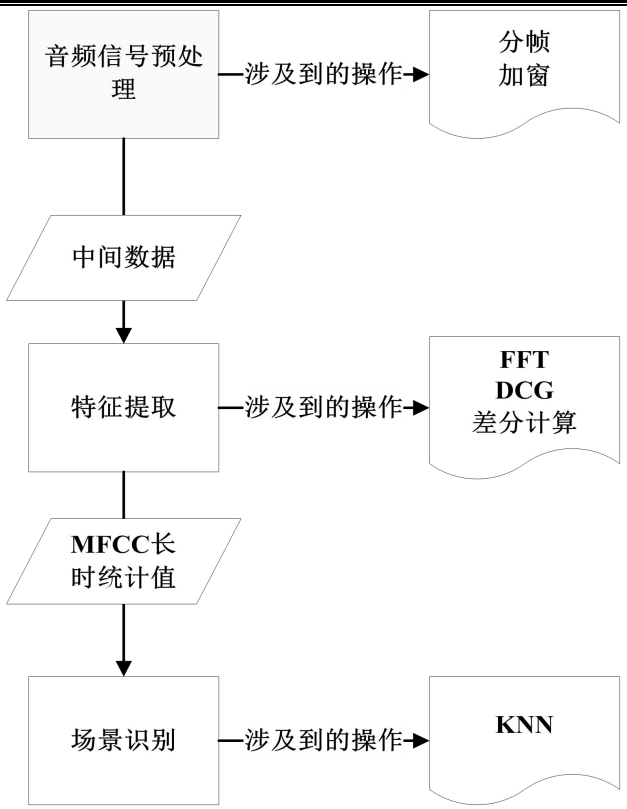

图 2-1 基线系统基本流程图

号，但是随着时间变化它的幅度值也在变化且变化较剧烈。现在成熟的信号处理技术往往分析的都是时不变信号即平稳信号，所以我们假设音频信号在短时间内具有平稳性，来适应传统的分析方法。为了得到上面假设的平稳信号，我们首先得对原始音频信号进行短时切分操作，即用窗函数平滑地在时间轴上滑动将整个信号分成以帧为单位的短时信号；不同的窗函数形式或者不同的窗长都会影响最终的结果。实验中我们采用 hamming 窗，计算公式如(2-1)所示，帧叠我们设置为窗长的一半。

$$W(n) = \begin{cases} 0.54-0.46*\cos[2n\pi/(N-1)], & 0\le n\le N-1 \\ 0, & else \end{cases} \tag{2-1}$$

上式中，N 为窗长。

2.2.2.2 特征提取

模式识别任务中，非常重要的一个环节就是提取训练样本数据的特征。特征数据的好坏会直接决定后面分类器分类结果的准确程度。

MFCC 具有其它声学特征不具备的性质，即它根据人耳的听觉机理对原始频率（Hz）进行了非线性映射，是在目前识别任务中最常用的音频特征,计算公式如(2-2)所示。

$$f_{Mel} = 2595 * \log(1 + \mathrm{f}/700) \tag{2-2}$$

上式中，f 为原始频率。MFCC 系数计算过程如下：

首先经过上面介绍的预处理操作（分帧、加窗处理）得到中间数据。

然后对上面得到的中间数据进行 FFT, FFT 是把音频信号从时域向频域进行转换，于是得到频谱数据，计算公式如(2-3)所示。

$$x_a(\mathrm{n}) = \sum_{k=0}^{255} x(k) * \mathrm{e}^{\frac{-j\pi ka}{128}}, 0 \le n \le 255 \tag{2-3}$$

上式子中 $x(k)$ 为中间数据，a 为帧序号。

对上面得到的频谱数据进行平方操作我们就得到能量谱，然后用 M 个梅尔带通滤波器对能量谱进行滤波操作，最后将所有经过滤波得到的子带能量进行叠加求和得到功率谱。

将上面得到的功率谱取对数，然后进行离散余弦变换,最后就得到 L 个我们要求的 MFCC 系数。一般 L 在 12 到 16 这一区间进行取值，实验中我们取 L 值为 12，计算公式如(2-4)所示。

$$C_n = \sum_{k=1}^{M} \log * x(\mathrm{k}) * \cos[\pi(\mathrm{k} - 0.5)\,\mathrm{n}/\mathrm{M}], \mathrm{n} = 1, 2, 3 \ldots .12 \tag{2-4}$$

上面得到的结果只是静态的音频特征，如果在静态音频特征基础上级联静态特征的差分参数对识别性能的提升会有很大的帮助，因为差分参数中含有动态音频特性。所以，实验中我们加入了一阶差分参数，最终得到的是 24 维的 MFCC 特征向量。

上述 MFCC 计算流程图如图 2-2 所示。

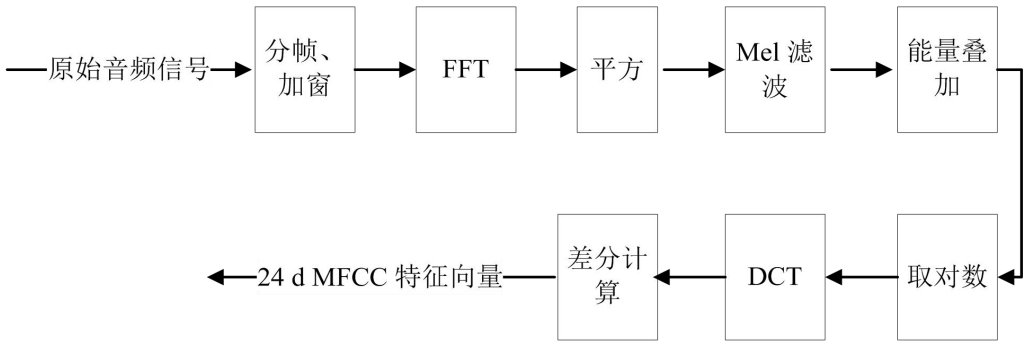

图 2-2 MFCC 计算流程图

2.3 场景识别

本课题，我们研究的是多类音频场景识别问题，选择的分类器是 k 近邻模

型。 给出训练数据集，如果要判断测试样本的类别首先得在训练集中找出与该样本最近的 k 个样本，然后统计这 k 个样本对应的所有类别中每个类别包含的样本数量，最后通过多数表决方式选出包含样本数量最多的类别做为测试样本的类别归属。虽然 KNN 不是目前最好的识别模型，但是我们的研究重心是音频场景的特征提取，而 KNN 具有简易、灵活和有效性，最终选择其作为场景识别的分类模型。下面我们对 k 近邻的理论基础和实验应用进行简单的介绍。

2.3.1 k 近邻算法的理论基础及算法流程

k 近邻算法既可以用来做分类又可以用来做回归，本课题使用 KNN 对音频场景进行识别属于分类问题。k 近邻算法的输入变量为样本的特征向量，对应为由特征向量构成的欧式空间中的点；输出结果则是和测试样例相对应的类别标签，而且输出结果可以是多个类别。

在已有的一个训练数据集上，k 近邻算法假设数据集中的样本实例的类别是已知的，在测试过程中对于新的测试样本，我们根据经过计算得到的与测试样例最近的 k 个训练样本所对应的类别，选出包含样本数量最多的类别作为测试样本的类别。因此，k 近邻算法没有其它分类器（svm、gmm 等）的显式训练过程。实际上我们可以把 k 近邻的分类过程看做是一个利用训练数据集对特征向量空间进行划分的过程[41]。

k 近邻算法中三个重要的基本因素为：k 值的选择、样本向量之间距离的度量方式以及最后的分类决策规则，下面我们对其进行简单的介绍，在此之前我么先说明下 k 近邻算法，如算法 2-1 所示。

算法 2-1: k 近邻算法

输入：训练样本集 $T = \{(x_1,y_1),(x_2,y_2),(x_3,y_3),....(x_n,y_n)\}$、k 值，其中 x_i 为特征向量，y_i 为对应样本的类别。

输出：测试样本 x_i 对应的 y_i。

首先：按照预先指定的距离计算方式，在训练样本集合 T 中找出与距离最近的 k 个样本点，用集合 $S(x_i,k)$ 表示。

然后：在 $S(x_i,k)$ 中按照预先指定的分类决策规则（如投票原则），得到测试样本 x_i 的类别 y_i，决策公式如下(2-5)所示。

$$y = \frac{\arg\max}{c_j} \sum_{x_i \in S(x,k)} I(y_i = c_j), i = 1,2....n; j = 1,2.....K \tag{2-5}$$

上式中 I 为指示函数，即当 $y_i = c_j$ 时，I =1；否则 I 为零。

2.3.2 KNN 模型及基本要素

2.3.2.1 KNN 决策过程

如果给定训练样本集合、样本之间的距离度量方式、分类决策规则还有 k 值，那么对于任意一个给定的测试样本实例，它的类别也就随之确定。决策过程如图 2-3 所示，在训练样本 x_i 对应的 $S(x_i,k)$ 中，我们根据投票原则找出包含样本数量最多的类别为圆圈，所以 x 的类别为圆圈。

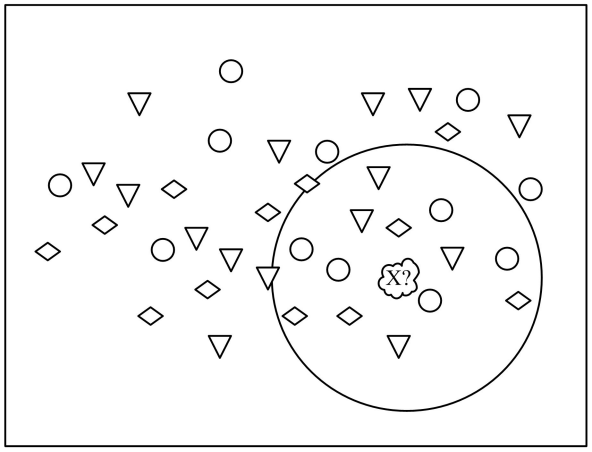

图 2-3 KNN 决策过程图

2.3.2.2 距离度量方式

训练样本特征向量之间的距离反映了样本之间的相似程度，不同的距离度量方式最终所确定的近邻点是不同的。常用的距离度量方式有欧式距离，曼哈顿距离等，我们可以用统一的闵可夫斯基距离又称为 L_p 范数(norm)来定义[42]，如公式(2-6)所示。

$$L_p(\mathrm{x}_i,\mathrm{x}_j)=\left\|\sum_{l=1}^{n}\left|x_i^l-x_j^l\right|^p\right\|^{\frac{1}{p}} \tag{2-6}$$

上式中，当 p=2 时，为欧氏距离，即：

$$L_2(\mathrm{x}_i,\mathrm{x}_j)=\left\|\sum_{l=1}^{n}\left|x_i^l-x_j^l\right|^2\right\|^{\frac{1}{2}} \tag{2-7}$$

当 p=1 时，为曼哈顿距离，即：

$$L(\mathrm{x}_i,\mathrm{x}_j)=\left\|\sum_{l=1}^{n}\left|x_i^l-x_j^l\right|\right\| \tag{2-8}$$

实验中，我们选择欧式距离作为样本特征向量之间的距离度量方式。

2.3.2.3 k 值的选择

不同的 k 值，导致最终的实验结果会有很大的不同。

如果 k 值选择过小的话，得到的近邻样本数量就会过少。其优点是最终的分类结果和近邻样本点类别分析的结果偏差变小，因为最终需要分析的决定测试样本类别归属的训练样本的数量相对变少。但是缺点是最终的分类结果和样本真实类别偏差会变大，即最终的分类结果对近邻样本点的依赖程度变大。如果近邻样本点为噪声数据的话，结果误差会很明显，即放大了噪声数据对分类结果的干扰作用，这就类似于过于复杂的模型出现过拟合现象[43-45]。

如果选择偏大的 k 值的话，结果就是近邻样本数量会偏多。这样的优点是分类的结果和真实类别的偏差变小，缺点是分类的结果和近邻样本类别分析的结果偏差变大。因为这种情况下近邻样本点中包含的噪声数据相对也会增多，距离测试样本点较远的训练样本点也会对最终的预测结果起一定的作用，类似于过于简单的模型出现欠拟合现象。

实际应用中，k 一般选择较小的数值且选择的 k 值一般不大于样本数量的平方根。

2.3.2.4 分类决策规则

k 近邻算法中的分类决策规则一般是投票原则。即由 $S(x_i,k)$ 中包含样本点多的类别来作为测试样本的类别。

如果假设分类的损失函数为 0-1 损失函数的话，分类函数如式(2-9)所示，那么投票表决规则就等同于经验风险最小化。原因如下：

$$f:R^n->\{\mathrm{c}_1,\mathrm{c}_2,....\mathrm{c}_k\} \tag{2-9}$$

首先误分类的概率如式(2-10)所示。

$$P(\mathrm{Y}\neq \mathrm{f(x)})=1-\mathrm{P}(\mathrm{Y}=\mathrm{f(x)}) \tag{2-10}$$

对于给定的测试样本实例 X，如果涵盖 S(x,k)的区域的类别是 c_j，则误分类率计算公式如(2-11)所示。

$$\frac{1}{k}\sum_{x_i\in S(\mathrm{x,k})} I(\mathrm{y}_i\neq \mathrm{c}_j)=1-\frac{1}{k}\sum_{x_i\in S(\mathrm{x,k})} I(\mathrm{y}_i=\mathrm{c}_j) \tag{2-11}$$

如果我们要使误分类数量最小，即经验风险最小，就要让 $\frac{1}{k}\sum_{x_i\in S(\mathrm{x,k})} I(\mathrm{y}_i=\mathrm{c}_j)$ 最大，所以投票表决规则等价于经验风险最小化。

2.3.2.5 k 近邻的 kd-tree 实现算法

我们在实现 k 近邻算法的过程中，最重要的环节就是找出 k 个近邻样本点，这就需要计算样本向量之间的距离。

最简单的计算方式就是遍历，即线性扫描。这样就要计算待测试样本特征向量和所有训练样本特征向量之间的距离，如果样本空间维度很高，样本数目很大的话，计算量非常的大，显然这种方式不可取。

kd 树是一种二叉树，它可以被看做是对 k 维空间的一个线性划分，这样我们就可以用它来进行存储 k 维样本数据点。

kd 树的构造过程是递归进行的，我们可以简单地描述其为递归的对 k 维空间进行切分，进而生成子结点，如果子结点内不再含有样本点就停止切分。递归过程中，我们会将相应的样本点保存在生成的子结点中。kd 树构造算法如算法 2-2 所示。

算法 2-2 kd 树构造算法

输入：k 维特征向量数据集 $S=\{\mathrm{x}_1,\mathrm{x}_2,\mathrm{x}_3,...\mathrm{x}_N\}$，其中

$\mathrm{x}_i=\{\mathrm{x}_i^{(1)},\mathrm{x}_i^{(2)},\mathrm{x}_i^{(3)},...\mathrm{x}_i^{(\mathrm{k})}\}^T, i=1,2,...N$

输出：构造好的 kd 树

首先：根节点的构造，从上面的分析我们可以看出根结点对应于包含 S 的 k 维空间的超矩形区域。

选择 $x^{(1)}$ 作为坐标轴，以 S 中所有样本实例点的坐标中的第一维数据的中位数作为切分点，以通过切分点并与坐标轴 $x^{(1)}$ 垂直的超平面为切分平面将根结点对应的超矩形区域进行切分，结果是得到两个子区域。这两个子区域分别对应为深度为 1 的左右两个子结点，这两个子结点的第一维坐标值分别小于和大于切分点在该维度上的值。最终落在切分超平面上的样本实例点保存在根节点上。

迭代：对深度为 h 的结点，选择第 l 维作为切分坐标轴，其中 $l=h(\mathrm{mod}\ k)+1$，以该结点对应的区域中所有样本实例点的第 l 维坐标值的中位数为切分点，以通过切分点并且与坐标轴 $x^{(l)}$ 垂直的超平面为切分面，把该点对应的超矩形区域进行切分，最终得到两个子区域；将落在切分面上的样本实例点存储在该结点上。

停止：按照上面两个步骤进行迭代，直到左右两个子节点对应的两个子区域不再含有样本点时停止，这时 kd 树的构造也就完成了。

K 近邻搜索是在上面构造完成的 kd 树上进行的，其相对于遍历搜索来说，可以省去大部分样本特征向量之间的距离计算，大大减少了计算量。

给定一个待测试样本点，要搜索其最近邻点。首先我们必须找到包含该实例点的叶子结点，然后以此叶子结点为出发点逐步向父结点回溯，在回溯过程

中我们要持续查找与目标样本点距离最近的结点，并做到随时更新当前的最近邻点；如果确定不可能存在更近的结点时停止搜索。如算法 2-3 为 kd 树的最近邻搜索算法。

算法 2-3 kd 树的最近邻搜索算法

输入：kd 树，目标样本点 x。

输出：x 的最近邻点。

首先：将目标样本点 x 从树的根节点开始查询，按照查询结果向下递归访问 kd 树，直至到达叶子结点位为止。

回溯：为了找到和目标样本点更近的训练样本点，从还未被访问过的分支里判断是否有离目标样本点更近的点。如果父节点下的其它分支里有更近的点，进入该分支，转向上一步并更新当前最近距离变量。回溯过程是从下往上进行的，如果回溯到根节点就说明已经不存在离目标样本点更近的点了。

上面的算法中判断未被访问过的分支中是否有离目标样本点更近的点时涉及到以下两个操作：

（1）如果该子区域中含有的实例点比当前最近距离小的话，那么我们就把它更新为最近邻点。

（2）当前最近邻点肯定内含于某个结点的其中一个子结点对应的矩形区域中，我们就检查其父结点的其它分支结点对应的矩形区域是否含有更近的点，如果没有就继续向上查找。

2.3.3 k 近邻算法应用

我们的样本数据是 24 维的特征向量，样本数目不足一万，所以即使是线性扫描，从计算量角度来说也是可以接受的。但是线性扫描实现更方便，所以实验中采用线性扫描进行邻近点的查询，计算流程图如图 2-4 所示 。

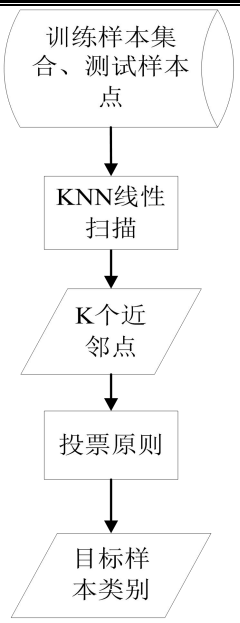

图 2-4 KNN 分类流程图

2.4 实验结果及分析

2.4.1 实验数据

实验中我们使用的场景语料库是由 17 个场景类别（机场、体育场、大街上、海滩、公交车站、教室、高速公路、超市等）共计 8298 个场景片段组成，每个场景片段为 2s 的音频段，其中同一类别的不同场景片段是在同一场景下不同位置或者不同时间段采集的。训练数据由 7617 个样例组成，剩余的 681 个样例作为测试样本用例。

因为本课题的研究重心是音频场景的特征分析，所以我们只是通过比较不同的音频场景特征在相同的识别模型上的识别率就可以了。

训练数据类别以及对应的样本数量详细信息如表 2-1 所示。

表 2-1 训练数据

场景类别	air	bas	bea	bus	cel	cla	cou	foo	hig	kit	mar	off	par
数量	246	298	375	678	500	407	300	323	420	503	552	177	481
场景类别	res	str	tra	Pro									
数量	347	717	683	610									

测试数据类别及对应的样本数量详细信息如表 2-2 所示。

表 2-2 测试数据

场景类别	air	bas	bea	bus	cel	cla	foo	hig	kit	mar	off	par	pro
样本数量	38	35	40	54	44	33	27	31	37	44	15	48	49
场景类别	str	tra	res	cou									
样本数量	67	62	36	21									

2.4.2 实验参数调整及结果分析

本课题的基线系统实验中主要调整的参数是 KNN 中的 k 值。调整策略是从较小值开始，逐步向增大、减小两个方向进行调整。

整体识别结果如图 2-4 所示，各个音频场景类别的识别结果与 k 值之间的关系如表 2-3 所示。

从表 2-3 可以看出场景类别中第 1、2、6、9、10、11、14 个场景的类别识别率在 k 取值为 2 时取得最高点，同时在场景类别为 5、12 的识别率都相对较高。整体识别率在 k 的取值为较小值时取得最佳效果，这也和图 2-4 中的识别率最高点相互对应。

2.5 本章小结

本章主要是围绕基于knn分类器和mfcc特征向量的音频场景识别基线系统进行了展开。首先给出了基线系统的整体流程，然后从数据预处理、特征提取、音频场景分类三个方面进行了介绍，详细的描述了各个环节的主要算法以及处理流程。通过调整影响最终分类性能的参数即 knn 中的 k 值，来进行音频场景识别性能的分析，最终得到基线系统的整体识别结果。

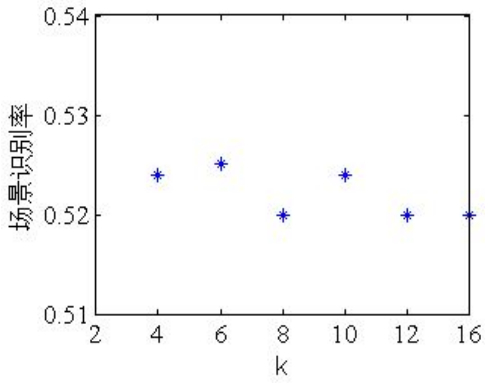

图 2-4 不同 k 值对应的场景识别结果

表 2-3 不同 k 值对应的场景类别识别结果

场景	K = 2	K = 4	K = 6	K = 8	K = 10	K = 12	K = 16
bea	0.80	0.78	0.75	0.75	0.75	0.75	0.75
air	0.35	0.27	0.29	0.29	0.29	0.29	0.32
bas	0.09	0.03	0.09	0.12	0.12	0.12	0.12
bus	0.84	0.84	0.84	0.81	0.80	0.84	0.86
cel	0.64	0.66	0.64	0.57	0.60	0.60	0.60
cla	0.79	0.76	0.73	0.67	0.66	0.67	0.67
cou	0.81	0.86	0.81	0.81	0.81	0.81	0.77
foo	0.56	0.60	0.63	0.63	0.63	0.63	0.67
hig	0.99	0.99	0.99	0.99	0.99	0.99	0.98
kit	0.82	0.82	0.79	0.82	0.81	0.79	0.78
mar	0.44	0.32	0.35	0.30	0.28	0.28	0.27
off	0.20	0.20	0.19	0.20	0.20	0.20	0.20
par	0.15	0.13	0.13	0.13	0.17	0.21	0.21
pro	0.13	0.11	0.11	0.11	0.13	0.11	0.09
res	0.14	0.14	0.17	0.17	0.20	0.20	0.20
str	0.72	0.77	0.74	0.72	0.71	0.72	0.72
tra	0.49	0.59	0.60	0.60	0.59	0.59	0.60

第 3 章 基于卷积神经网络的音频特征提取及场景识别

3.1 引言

现有的音频特征往往是从时域或者频域进行分析的。时域特征分析我们可以从短时能量分析、过零率分析等角度进行；频域特征分析可以从滤波器组分析、傅里叶频谱分析等角度进行；此外音频特征还有基于人类听觉特性同时也是当前应用最广泛的梅尔频率倒谱系数。MFCC 是基于频谱图中共振峰的位置及其变化轨迹提取出的频谱包络，它对于结构性较强的音频信号（说话声、音乐等）有着很好的描述刻画能力。但是音频场景信号是自然音频信号，频率变化较剧烈，而且还有很多背景噪声。如果使用短时特征(MFCC)，就不能完整的刻画出音频场景的声学特征；如果使用长时统计值的话，会造成特征的局部结构性信息的丢失，最终都会导致场景识别性能下降。

时域信号和频域信号都是一维信号，如果我们联合这两者进行分析，把时间和频率同时作为自变量，而把对应的能量值当做因变量，这就形成了时间频率平面上的二维信号。我们可以根据它来计算某一特定时间的频率密度，重要的是它对结构性较差的音频场景数据保留了时间、频率两个维度的信息，相比较 MFCC 而言，其对于音频场景识别会有更好的效果。所以，本章节我们使用卷积神经网络(Convolution Neural Network, CNN)对音频场景进行时频特征分析，提取能够反映结构性的长时特征来进行音频场景识别。

3.2 现有音频特征在场景识别方面的适用性分析

基线系统选择的音频场景识别的声学特征是 MFCC，所以这一小节我们对使用 MFCC 和 MFCC 的长时统计值进行音频场景识别存在的不足进行量化分析。

MFCC 是以帧为单位进行计算得到的，而帧长一般都选择 30 毫秒左右，这说明 MFCC 这种短时特征只是对很小的音频单位进行分析。然而，我们的音频场景信号是 2 秒的音频段，其中内含的关键声学事件的时长也大都为几百毫秒，所以，如果使用 MFCC 来进行音频场景识别的话，显然不能够完整的对音频场景的信息进行刻画。

MFCC 长时统计值是通过对 MFCC 在较长一段时间内求均值得到的。海滩

场景信号一般主要由浪花拍打沙滩的声音和海鸥的叫声等关键声学事件组成，其中两种声学事件发生的绝对时间和相对时间都不确定，如果用 MFCC 长时统计值的话,相同的场景得到的统计值结果差别可能会很大，造成误识；同样，海滩和马路两个场景内含的关键声学事件的时长都在 200 毫秒左右，如果使用 MFCC 长时统计值的话，这两个场景对应的结果会很相似，也会造成误识。

3.3 基于 CNN 的深层特征提取原理及学习算法

CNN 通过对输入数据的逐层非线性处理，能够得到对非常复杂的数据分布更加抽象更加有效的描述信息，进而从中得到数据的深层特征。

CNN 在对语谱图进行分析时，其卷积滤波器是对局部时间上若干子带进行分析的，这样会保留语谱图的局部结构性，而滤波器在整个语谱图上进行遍历，最终覆盖整个语谱图，实现了对整个时长信号进行综合分析。

基于此，实验系统尝试把 CNN 在图像特征提取方面的一些技巧应用在对语谱图的处理上，来对音频场景信号进行深层特征分析，最后利用提取到的深层特征进行音频场景的分类。

本小节主要介绍 CNN 的理论知识。首先通过一个简单的拓扑结构图对其进行直观意义上的说明；然后通过其特征图的计算方式，描述了在给定收敛的网络模型前提下进行输入数据特征值的计算过程；最后给出了网络模型的训练机制。

3.3.1 CNN 拓扑结构

如图 3-1 所示为一个卷积神经网络的拓扑结构示意图，除了输入输出层外，每层由多个二维平面组成，每个平面由多个相互独立的神经单元构成。

从下面的拓扑结构示意图可以看出，CNN 从左由输入层开始向右依次连接卷积层和下采样层，其中 C 表示卷积层 S 表示下采样层。卷积层和下采样层可以设置多个且它们之间是间隔设置的，最后按照全连接方式连接几个普通网络层。CNN 的输入数据为二维矩阵，在本实验系统中，输入数据就是语谱图。

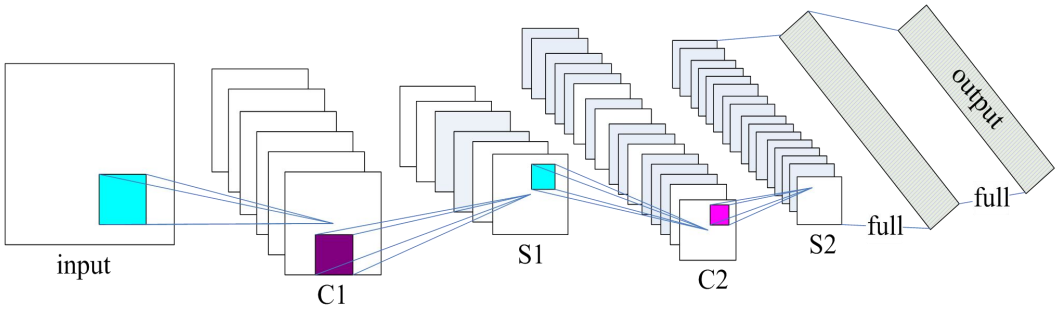

图 3-1 卷积神经网络拓扑结构示意图

3.3.2 CNN 计算方式

这一小节介绍给定一个卷积神经网络模型，如何进行特征图的求解计算问题，即卷积神经网络的前向传播过程。前向传播过程中唯一的一个常量是输入矩阵，其它像卷积滤波器对应的矩阵系数、偏置系数以及最后的全连接层中的矩阵系数都是随机初始化的较小值，它们都是在后面的学习过程中待学习的参数变量。

把输入矩阵记为 input，其大小记为：input_x × input_y；卷积滤波器用 filter 表示，不同层次的卷积滤波器的大小会有所不同，统一记为：filter_x ×filter_y；下采样记为：scale，不同层次的下采样尺寸也会有所不同，也统一记为：scale_x × scale_y；最后是特征图矩阵的大小，特征图矩阵这里用其名字加下划线的形式表示，比如第一个卷积层特征图矩阵就用 C1_表示，它的大小就用：C1_x × C1_y 来表示。

第 1 个卷积层 C1 的特征图（由 n1 个特征图组成）中的特征单元值是由输入矩阵和卷积滤波器进行卷积运算得到的，计算公式如(3-1)所示，其中 act_C1_i 表示特征图中第 i 个特征单元值，f 是为了得到输出值而用到的激活函数（sigmoid、tanh 等），它起到非线性变换的作用。这样，C1 特征图的大小为：(input_x-filter_x+1)×(input_y-filter_y+1)。

$$act\_C1_i = f(\text{input}*\text{filter}_i),\ i = 1,2,3....n1 \tag{3-1}$$

上式中*为卷积运算，可以看出第个特征单元的值是由输入矩阵和对应的第个卷积滤波器进行卷积之后再进行非线性换换得到。

第 1 个下采样层 S1(特征图数量和 C1 相对应，也是 n1）对应的下采样过程如下：把 C1 特征图中相邻尺寸为 scale_x×scale_y 大小的二维区域看做一个处理单元，然后对该单元做最大池化（max pooling）处理，对应的特征单元值的计算公式如(3-2)所示；或者对该单元做平均池化（average pooling）处理，对

应的特征单元值的计算公式如(3-3)所示。其中 act_S1_i 表示 S1 特征图中的第 i 个特征单元，max 表示取最大值操作，avg 表示取平均值操作。S1 中特征图的大小为：(C1_x/scale_x) * (C1_y/scale_y)。

$$act\_S1_i=max(act\_C1_{i,j}),\ j \in act\_C1(scale\_x \times scale\_y) \tag{3-2}$$

$$act\_S1_i=avg(sum(act\_C1_{i,j})),\ j \in act\_C1(scale\_x \times scale\_y) \tag{3-3}$$

第 2 个卷积层 C2（n2 个特征图）的计算方式有所不同，这是因为 S1 由多个（n1 个）特征图组成，故 S1 和 C2 之间的连接方式不同于普通的层间连接。假定 n1=6, n2=16, 则 S1 和 C2 之间连接如表 3-1 所示（+表示特征图之间有连接）。

表 3-1 池化层和卷积层之间的连接方式

C2 S1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	+				+	+	+			+	+	+	+		+	+
2	+	+				+	+	+			+	+	+	+		+
3	+	+	+				+	+	+			+		+	+	+
4		+	+	+			+	+	+	+			+		+	+
5			+	+	+			+	+	+	+		+	+		+
6				+	+	+			+	+	+	+		+	+	+

C2 中的特征单元值的计算方式同上面的计算方式一样，由 S1 中对应的特征图和对应的 filter 进行卷积求和得到。特征图大小是：(S1_train_x-filter_x+1)×(S1_train_y-filter_y+1)。

第 2 个下采样层 S2 含有的特征图数量和 C2 是相互对应的。和上面一样，还是用最大池化操作或者平均池化操作来计算该层的特征单元值,其中该层的特征图大小为：(C2_x/scale_x)× (C2_y/scale_y)。

S2 计算完成后，继续向右进行其它网络层次的设置。或者根据需要继续进行卷积层和下采样层的间隔设置然后再进行全连接层和输出层的设置，后续或者直接进行全连接层和输出层的设置。全连接层和输出层如图 3-1 所示, S2 右边为全连接层，继续向右为输出层，输出层神经单元数量对应为音频场景类别的数量。

3.3.3 CNN 学习算法

上面介绍的 CNN 的计算过程对应为网络的前向传播过程，本小节将要介绍的 CNN 的学习算法是用来对网络中的参数的进行更新的，该更新过程对应

为网络的反向传播过程。

以 3.3.1 中的简单拓扑结构图对应的 CNN 为例，进行学习算法的描述。

输入数据：固定尺寸的二维矩阵 $input$ 以及对应的标签 Y 。

待学习的参数：卷积层对应的卷积滤波器的权重矩阵 w 和偏置 b ，下采样层特种图神经单元对应的乘性偏置 β 和特征图对应的线性偏置 b 。

损失函数：实验中取平方损失即 $L(input,Y) = \sum(Y - f(input))^2$ ，其中 $f(input)$ 表示网络最终的输出值。

我们记每个特征图中的特征值为 z ,经过激活处理的输出值为： a ，卷积滤波器用 $filter$ 表示。

从上面的拓扑结构图和计算方式我们可以看出：卷积层和下采样层神经单元值的计算方式是不同的，所以对应的参数更新过程也是有区别的，这样 CNN 的学习算法和普通的 BP 神经网络的学习算法也就不会相同。下面，我们对 CNN 的学习算法进行逐层独立分析。

（1）卷积层

首先通过和卷积层相对应的输入层包含的特征图和一个待学习的卷积滤波器进行的卷积运算，然后再通过激活函数的非线性映射，我们就得到了卷积层特征图。卷积层的每一个特征图可能是由多个输入层的特征图通过卷积运算得来(图 3-1 中 S1 和 C2)的。所以卷积层中神经单元值的计算如公式(3-4)所示。

$$a_j^l = f(\sum_{i \in M_j} a_i^{l-1} * filter_{i,j}^l + b_j^l) \tag{3-4}$$

上式中 M_j 表示卷积层的输入层中所有待卷积的特征图集合， l 为网络层次指示变量， a_j^l 表示第 l 层中的第个 j 输出单元值， b_j^l 表示和 a_l^j 对应的加性偏置。下面是特征值和输出值的层间递推计算，如公式(3-5)、(3-6)、(3-7)所示。

$$l = 0.5 \times \sum_i (a_i - y_i)^2 \tag{3-5}$$

$$z^l = w^l \times a^{l-1} + b^l \tag{3-6}$$

$$a^l = f(z^l) \tag{3-7}$$

其中(3-5)中的 a_i 表示输出层第 i 维的输出值， y_i 表示类别向量的第 i 维分量。

如果我们定义： $\delta = \frac{\partial l}{\partial z}$ ，则权重矩阵 w 和偏置 b 的梯度的计算如公式(3-8)、(3-9)所示。

$$\frac{\partial l}{\partial b} = \frac{\partial l}{\partial z} \times \frac{\partial z}{\partial b} = \delta \tag{3-8}$$

$$\frac{\partial l}{\partial w^l} = \frac{\partial l}{\partial z} \times \frac{\partial z}{\partial w^l} = a^{l-1} \times (\delta^l)^T \tag{3-9}$$

上式中δ的递推公式如(3-10)所示。

$$\delta^l = (w^{l+1}) \times \delta^{l+1} \times f^{'} \tag{3-10}$$

上式中可以看出，第l层的δ值就等于上面一层的δ值和对应的权重矩阵的转置的乘积，然后再乘上对应的激活函数的导数。

从上面的公式可以看出，想要更新权重矩阵参数，必须求出δ^l，这就要求先求出$\delta^{(l-1)}$。从上面的卷积神经网络的拓扑图可知，下采样层特征图中每个特征单元值的计算，是通过卷积层特征图中采样窗口大小的区域而不是特征图的全部进行的，所以卷积神经网络中的δ^l的计算公式有如下改变。

$$\delta_j^l = w_j^{l+1} \times (f^{'} \times \mathrm{up}(\delta_j^{l+1})) \tag{3-11}$$

上式中 up 为上采样操作，即沿着水平垂直两个方向对单元值进行拷贝，这样我们就会得到一个矩形区域，其值等于单元值，最终对应的矩形区域大小和采样窗口大小相同。卷积神经网络中的变量δ的初始值计算公式如 (3-12)所示。

$$\delta^l = f^{'}(\mathrm{z}_i^l) \times \sum_i (a_i^l - y_i) \tag{3-12}$$

上式中，i为输出层特征向量维数的指示值。

这样，对于一个给定的卷积层特征图，梯度计算公式如(3-13)、(3-14)所示。

$$\frac{\partial l}{\partial b_j} = \sum_{u,v} (\delta_j^l)_{uv} \tag{3-13}$$

$$\frac{\partial l}{\partial w_{i,j}^l} = \sum_{u,v} (\delta_j^l)_{uv} (p_i^{l-1})_{uv} \tag{3-14}$$

上式中$(p_i^{l-1})_{uv}$对应为a_i^{l-1}中与卷积滤波器进行卷积运算的局部感受野，u 和 v 是局部感受野水平和垂直两个方向的指示变量。

（2）下采样层

重构误差关于乘性偏置和线性偏置的梯度计算公式如(3-15)、(3-16)所示。

$$\frac{\partial l}{\partial b_j} = \sum_{u,v} (\delta_j^l)_{uv} \tag{3-15}$$

$$\frac{\partial l}{\partial \beta_j} = \sum_{u,v} (\delta_j^l \times down(a_j^{l-1}))_{u,v} \tag{3-16}$$

上式中，$down()$是对第$l-1$层的采样窗口区域进行下采样操作。

3.4 基于 CNN 的场景特征提取及识别分析

上面我们分析了在场景识别任务中目前已有的音频特征的适用性，并指出了时域、频域特征的不足之处。而时频分析作为时变非平稳信号分析的有力工具，能够给出各个时刻的瞬时频率及其幅值，能够为我们提供时间和频率两个方向的联合分布信息。本小节对 CNN 如何能够进行音频场景数据的特征提取进行理论分析，并给出最终的音频场景识别算法。

3.4.1 语谱图的特性以及 CNN 方法的适用性

时频分析我们可以从短时傅里叶变换和小波变换等角度进行。本课题实验系统用到的时频分析数据-语谱图是通过短时傅里叶变换得到的，下面是语谱图的计算过程。

首先进行分帧加窗处理，窗函数这里选择 hamming 窗，分帧过程中帧叠设置为窗长一半。这样原来的时序信号$x(\mathrm{n})$就可以表示成$x_n(\mathrm{m})$，其中变量 n 表示帧指示序号，变量 m 表示对应帧内的时间指示序号。

然后进行离散傅里叶变换（Discrete Fourier Transform，DFT），可以得到短时幅度谱数据。计算公式如(3-17)所示，其中$X(\mathrm{n},\mathrm{w})$就是$x(\mathrm{n})$的短时幅度谱估计。

$$X(\mathrm{n},\mathrm{w}) = \sum_m x_n(\mathrm{m}) \times e^{\frac{-2m\pi jw}{N}} \tag{3-17}$$

这样，时间点 m 处的频谱能量密度函数$P(\mathrm{n},\mathrm{w})$可以按照公式(3-18)进行计算，其中 conj 为取共轭操作。

$$p(\mathrm{n},\mathrm{w}) = \left|X(\mathrm{n},\mathrm{w})\right|^2 = (X(\mathrm{n},\mathrm{w})) \times (\mathrm{conj}(X(\mathrm{n},\mathrm{w}))) \tag{3-18}$$

将$p(\mathrm{n},\mathrm{w})$的值表示成灰度级所构成的二维图像就可以得到语谱图，通过变换：$10*\log_{10}(p(\mathrm{n},\mathrm{w}))$就可以得到语谱图的 db 表示。可以看出虽然语谱图是二维的图像，但表示的却是三维的信息。图像上面颜色的深浅表示了对应时间和频率处的能量大小，如图 3-2 所示（横轴表示时间，纵轴表示频率）。

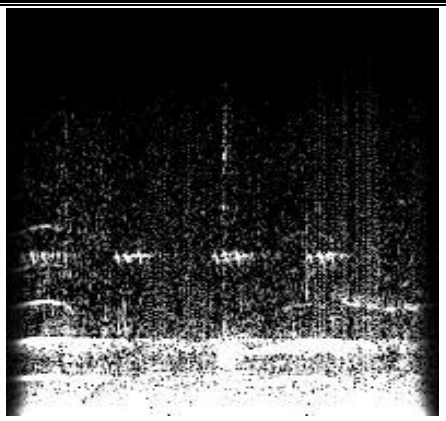

图 3-2 特定场景语谱图

选定语谱图作为我们分析的对象，是因为语谱图上体现出了两个方向的信息，但是这也要求我们具备相应的具有三维信息分析能力的方法，卷积神经网络以它独有的性质正好符合我们的要求。下面是 CNN 对语谱图进行深层特征分析的理论基础。

局部感受野：提取局部特征。普通的神经网络每次分析的对象都是整个输入变量，而我们在语谱图中进行分析的时候需要关注的是局部特征，所以我们分析的对象只能是语谱图中的一部分。卷积神经网络的卷积操作正好是通过卷积滤波器和语谱图中一小块区域的卷积运算进行的，具体为卷积滤波器沿着时域方向和频域方向有重叠地向前移动，每次分析的对象为具有一定中心频率的子带在长时间段内的局部信息。相当于时频滤波一样，这会保留下较长时间段内的频率方向的信息，同时还有增强局部特征、降低噪声的能力。如图 3-3 所示，上半部分为海滩场景的三个场景语谱图，下半部分为对应的三个卷积层特征图（第一个卷积层）。可以看出，低频噪声部分的影响在第一个卷积层中显然被降低了，同时局部有效内容也被加强了。

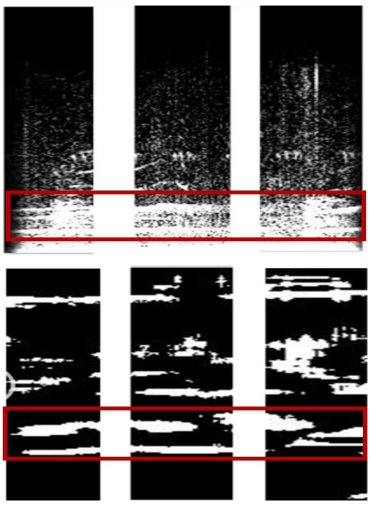

图 3-3 卷积处理效果图

下采样：下采样是通过对较小局部信息进行有效的抽取保留操作进行的，它会使 CNN 对信号中的有效内容的位移现象更加鲁棒。例如如果体育场场景语谱图中对应篮球撞击地面的声音段无论何时出现，其对应的场景类别都为篮球场，出现的早与晚不那么重要。这种混淆位置的处理方式对信号的变形或扭曲等现象有较强的抗畸变能力。此外，下采样操作在保留有用信息的同时还有效减弱了其它无关信息的干扰，比如说像是体育场场景中篮球撞击地面间隔的语音段在下采样的过程中就会被逐渐忽略掉。

从上面的分析中我们能够看出，CNN 在卷积和下采样运算过程中不仅会保留下含有频率信息的长时结构性特征，而且会减弱噪声的影响，同时对有效音频段在时间轴上的位置产生了一定的抗畸变能力。和之前的短时特征以及长时统计特征值相比这种长时结构性特征能够更好的反映出音频场景声音段中相对时长较长的有效内容，进而能够更好地进行音频场景识别。

3.4.2 卷积滤波器的设计

CNN 用于音频场景特征提取很重要的一个环节是卷积滤波器大小的设计。上面我们已经提到过卷积滤波器是用来进行长时局部特征分析的，所以在时域方向卷积滤波器的尺寸一般设计为与有效内容时长相当，频率轴方向的尺寸一般设计为频率子带宽度大小。这样卷积操作得到的特征图中的元素就对应了长时子带的特征。

我们把整个频带均匀划分为 6 个子带，初步设定为卷积滤波器在频率轴方向的尺寸为每个子带的宽度。通过我们的人耳分析，大部分的音频场景语料中的有效内容的时长为整个语音段时长的 1/8 左右，所以我们把卷积滤波器在时域方向的尺寸设置为整个时长的 1/8，如图 3-4 中红色区域所示。

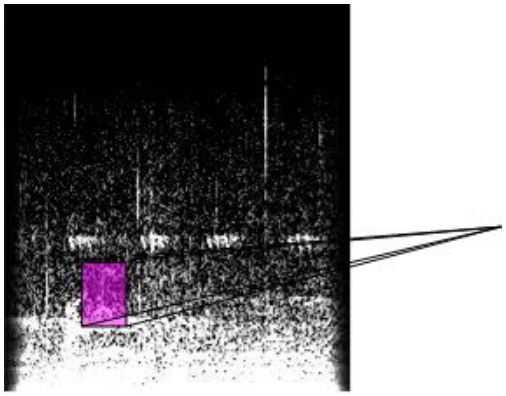

图 3-4 卷积核设计方案示意图

3.4.3 特征提取及分类

上面已经讲述了卷积滤波器的设计原则以及其在实验系统中的尺寸大小。接下来是音频场景特征提取以及最终的音频场景分类过程。

卷积神经网络的训练和分类是同时进行的，即它是一个端对端的生成模型。这样进行的好处是可以通过自学习进行自适应调节，主动学习找到最佳分类特征而不是过多地依赖主观经验。下面是特征提取及最终的分类算法：

（1）音频信号预处理得到时频联合分析数据；

（2）通过卷积层的卷积运算得到初步特征矩阵；

（3）通过下采样层的下采样运算得到最终的特征矩阵；

（4）训练最终的分类器即分类层和输出层之间的权重参数，进行最终的音频场景的分类。

其中，训练过程中我们采用经典的最优化算法批量梯度下降法[46,47]进行参数的更新，如果学习速率这个参数难于调整的话，还可以尝试使用牛顿法或拟牛顿法等优化算法[46,47]。

整个过程中，第一步进行的是最终的分类层参数的训练，首先把输出场景类别和真实场景类别之间的差值作为损失函数值，然后求出损失函数关于全连接层权重矩阵的梯度；第二步进行的是下采样层的参数训练，即按照层间梯度递推公式进行更新；最后进行的是卷积层的参数训练，这里需要一个上采样操作；网络的全部参数更新一遍记为一次迭代。如此经过若干次迭代得到最终收敛的卷积神经网络模型。

在上面得到的收敛的网络模型上进行音频场景的特征提取，然后进行类别预测，最后统计得到最终的音频场景识别率。

3.5 实验过程及结果分析

3.5.1 实验数据

音频场景语料和第二章相同，共计：7650 个训练样本和 648 个测试样本，17 个音频场景（airport、beach、bus station、classroom、street、highway 等）。每个样本数据都是 2second 的音频片段，采样频率为 22050Hz,采样点为 16。

按照内容 3.4.1 所述把音频时域信号转换到时频平面上进行分析，最终得到的语谱图图像的像素尺寸大小为：198×221，如图 3-3 所示。

3.5.2 实验基本网络结构

完成上面的数据预处理之后，现在分析的对象已经从一维音频数据转换到二维矩阵数据，这样场景识别问题就转换为类似图像分类问题。

实验中我们设计的网络结构详细如下：输入层为 221×198 的二维矩阵，然后是卷积层和下采样层重复交叠设置两次；最后以全连接方式连接一个输出层进行场景的分类，整体结构为六层网络结构；其中最后的输出层对应到 17 个场景类别，整体拓扑结构如下图 3-5 所示。

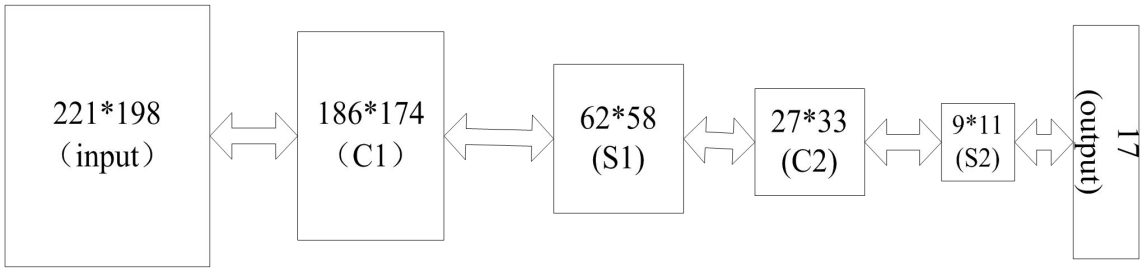

图 3-5 CNN 拓扑结构示意图

3.5.3 实验参数调整及结果分析

3.5.3.1 卷积滤波器大小调整及结果分析

从上面卷积神经网络的学习方法我们可以看到，卷积滤波器大小反映了每次待卷积处理的局部区域（局部感受野）的大小，所以会直接决定所提取的局部特征的粒度。矩阵大，对应的特征粒度就大，反之特征粒度就小。如果卷积滤波器大小和待分析的矩阵的尺寸一样的话，分析的就是整个对象，特征粒度就对应为全局了。

下面从特征粒度增大和减小两个方向进行参数的调整，具体调整内容如表 3-2 所示。

实验过程中，固定其它变量，只是按照上面的方案进行粒度大小的调整，迭代 20 次的结果如图 3-6 所示。

表 3-2 卷积核调整方案内容

方案 调整 size	Original	Case1	Case2	Case3	Case4
C1	36*25	66*55	33*22	39*28	18*13
S1	Scale = 3	Scale = 3	Scale = 3	Scale = 3	Scale = 3
C2	36*26	26*16	34*27	38*31	21*19
S2	Scale = 3	Scale = 3	Scale = 3	Scale = 3	Scale = 4

上表说明：case1 和 case3 是向着卷积滤波器尺寸增大的方向即特征粒度增大方向调整，case2 和 case4 是向着特征粒度减小方向调整，从结果图中我们可以看出特征粒度减小方向的识别结果都好于粒度增大方向的结果。结果说明卷积滤波器越小即特征粒度越小，分析的局部特征越细致；相反特征粒度越大，局部特征分析越粗糙分析能力也会随之下降。

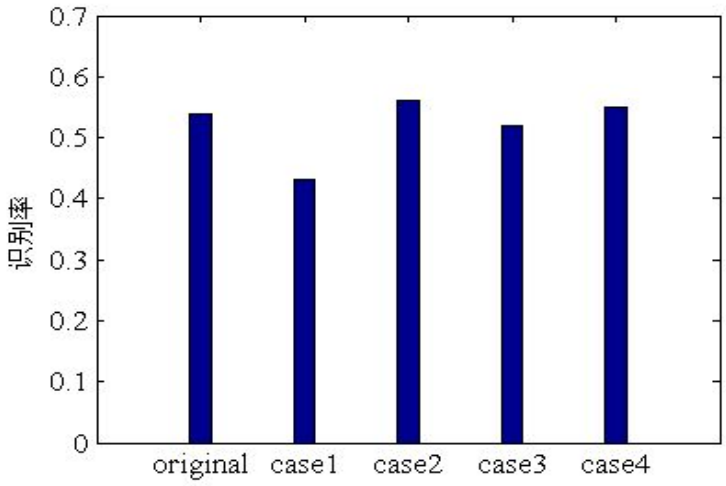

图 3-6 卷积滤波器调整对应识别结果图

3.5.3.2 特征图数量的调整及结果分析

从上面的卷积神经网络的学习方式可以看出，不同的权重矩阵对应不同的特征图，即从不同角度进行特征的分析,所以改变特征图的数量就会改变特征分析的角度。如果我们增加特征图的数量就可以增加特征分析的角度，这样一来就会增加特征分析的完备性，但是弊端是难免会造成冗余，而且计算量也会随之增加；相反，如果我们减少特征图的数量则会造成网络特征分析的角度减少，特征分析能力也会随之变得相对单薄。所以，下面对特征图数量这一参数从增加和减少两个方向进行调整，进而找到最佳数量点。

如表 3-3 所示是调整策略内容。

表 3-3 特征图数量调整策略内容

方案 特征图数量	Original	Case1	Case2
C1	4	7	3
S1	4	7	3
C2	8	14	6
S2	8	14	6

实验过程中，固定其它变量，按照上面的调整方案迭代 20 次的结果如图 3-7 所示，可以看出增加特征图数量识别性能有所提升，减少特征图的数量性能有所下降。实验结果和上面理论分析的相吻合，即增加特征图数量会使特征分析结果更加的完备；相反减少特征图的数量会让特征分析显得比较单薄，造成识别性能下降。

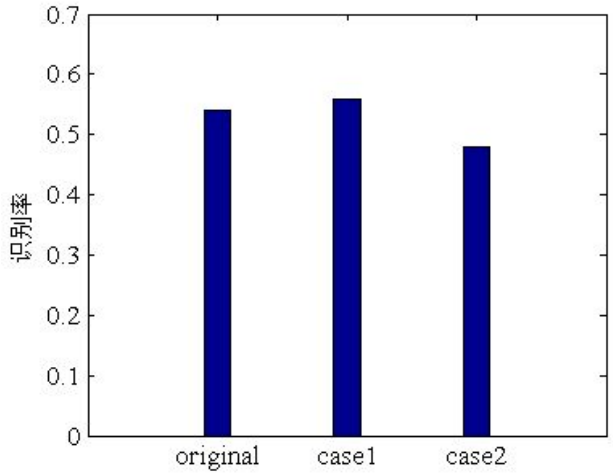

图 3-7 特征图数量调整对应识别结果图

3.5.3.3 激活函数参数调整及结果分析

神经网络中的激活函数主要是对隐单元进行非线性映射的，所以激活函数往往具备某些性质：首先也是必要的性质就是非线性、此外还有非必要性质如单调性、连续性、光滑性等[41]。

理论上允许神经网络的不同隐层可以有不同形式的激活函数，甚至同一层的不同神经单元也可以对应到不同的激活函数。我们实验中按照常规思路，激活函数是唯一的。常用到的激活函数有：sigmoid、tanh,此外还有最近被提出的像 relu、Lrelu 还有 Prelu 等。此外，对于不同的应用场景，同一激活函数所带来的性能也会不同，甚至有很大的区别。下面，首先简单介绍下实验中应用到的激活函数：

(1) sigmoid：目前使用较为广泛，产生的结果就是把隐单元的值映射到 0 和 1 之间，如图 3-8 所示。

该激活函数通常会导致更少的隐单元被激活，进而得到一个局部表示。计算公式如(3-19)所示。

$$f(\mathrm{x}) = \frac{1}{1+\exp(-\mathrm{x})} \tag{3-19}$$

上式中 x 为隐层单元的输入。

(2) tanh：双曲正切弯折，它可以把隐单元的值映射到-1 和 1 之间,如图 3-9 所示。和 sigmoid 相比，它更加鲁棒，但是缺点是对梯度消失更加敏感，导致收敛速度慢，甚至可能收敛到局部最优值。

(3) Rectified Linear Unit：相比较于 sigmoid 而言，往往能抽象出更稀疏和更离散的特征。特征分析往往会追求稀疏性，但是仅仅满足稀疏性也是不够的，因为它不能完全的反映出编码的质量（如果方差太大，导致熵太小，结果也不会太好）。

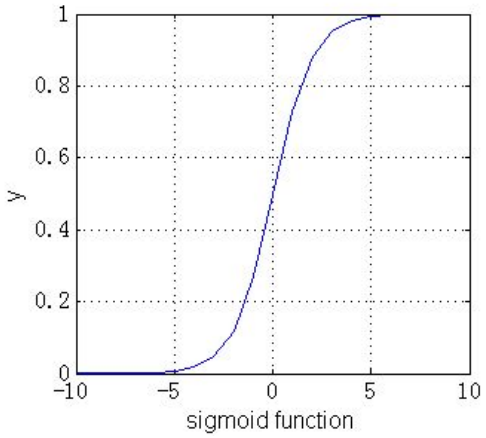

图 3-8 sigmoid 激活函数

离散性能够反映出激活单元对每一个激励所表现出的差异性（即熵较大），Relu 映射关系如图 3-10 所示，计算公式如(3-20)所示，它正是从稀疏性和离散性两方面进行考虑的，所以往往能够带来更好的性能。

$$f(\mathrm{x}) = max(x, 0) = \begin{cases} x \ ; x > 0 \\ 0 \ ; else \end{cases} \tag{3-20}$$

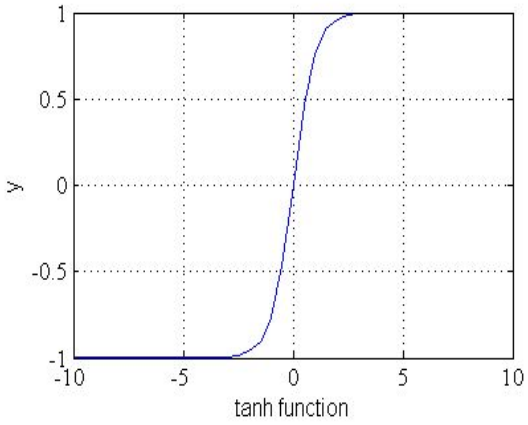

图 3-9 tanh 激活函数

(4) leak relu: 和上面的 relu 区别不大，就是在变量小于 0 的情况下做了 leak

处理公式如(3-21)，训练时的效果往往表现为收敛的较快。

$$f(x) = max\_leak(x,0) = \begin{cases} x\ ;x > 0 \\ a \times x\ ;else \end{cases} \tag{3-21}$$

上式中的a为一较小的常数 max_leak 表示进行弱化的取最大值操作。

(5) Parametric rectified linear unit :和 Lrelu 相比不同的是 Lrelu 中输入变量小于零时，对应的输出值为输入变量的常数倍，且常数往往取很小的数值。而 Prelu 在输入变量小于零时，输出值为输入变量的可变参数倍，计算公式如(3-22)所示。

$$f(x_i) = max(x,0) = \begin{cases} x_i\ ;if(x_i > 0) \\ a_i \times x_i\ ;else \end{cases} \tag{3-22}$$

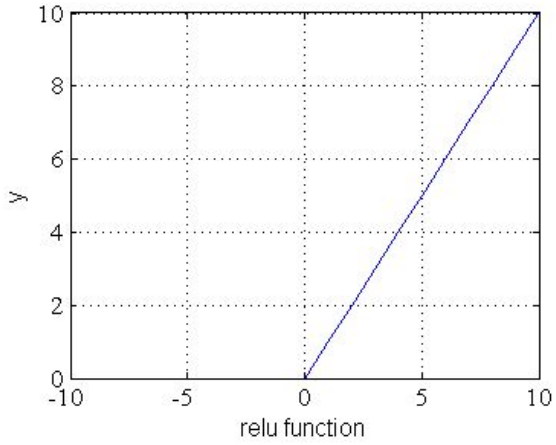

图 3-10 relu 激活函数

从上式可以看出，当输入矢量的维度分量小于零时，对应到每个单元都有一个可调参数，该参数和网络层间的权重参数一块作为整体参数进行更新。

实验中我们固定其它变量，按照上面说明的各个激活函数的处理方式进行改变激活函数的形式，实验结果如图 3-11 所示。可以看出，在音频场景识别这项识别任务中，sigmoid 的识别效果明显好于其它几个激活函数。

3.5.4 实验流程及中间结果分析

输入层为二维语谱图，对应的数据是 221×198 的二维矩阵。向前到第一个卷积层，对应的卷积滤波器大小设置为：36×25。该卷积滤波器所分析的二维对象在时间轴上为时长大约为 250ms 的音段,频率轴上大约为频带宽度六分之一的子带。卷积滤波器进行卷积运算时向前移动的步幅为 1，实验的中间结果如图 3-12 所示。

从下图的中间结果可以看出卷积层四个特征图的内容各不相同，且相差较

大。说明对输入层的进行刻画描述的角度不同，最外面的特征图对输入图像的轮廓描述的较为清晰。

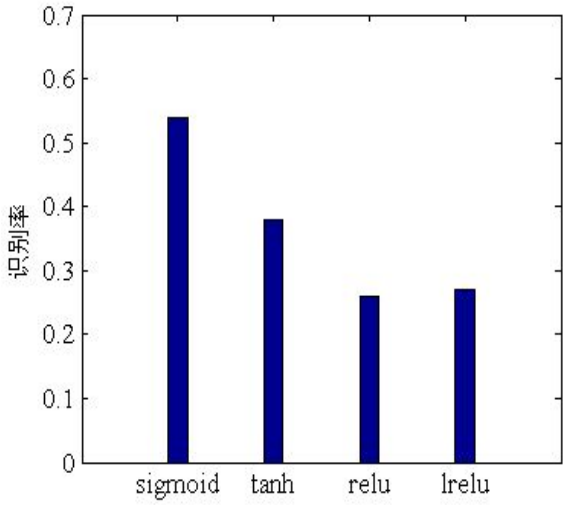

图 3-11 激活函数调整对应识别结果图

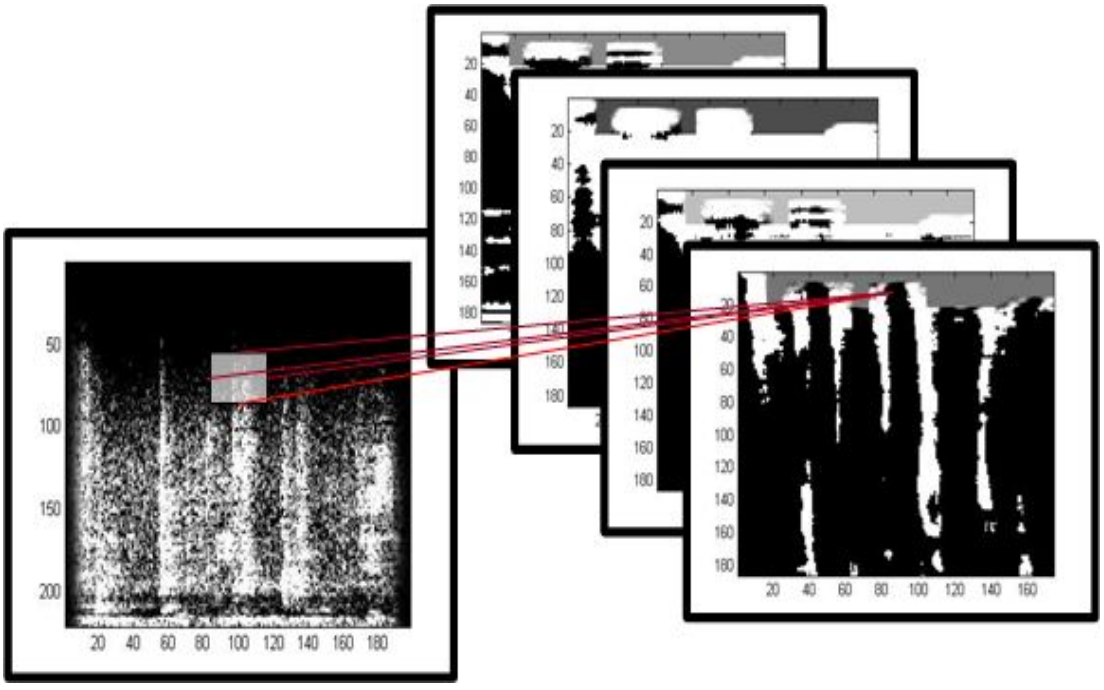

图 3-12 输入层-卷积层

第一个卷积层中共有 4 个特征图如上图所示，继续向前到第一个下采样层，对应的采样窗口大小设置为：3×3。即在卷积层特征图上大小 3×3 的二维窗口里共计 9 个特征点中提取出一个代表元素最为该窗口的特征，按照这种方式有重叠地把整个特征图遍历一遍，最终得到具有一定抗畸变能力的特征图。实验

的中间结果如图 3-13 所示。

从下面的中间结果图中我们可以看出，下采样层中四个特征图内容和上层对应的卷积层的内容很相似。这和我们上面分析的是一致的，整体效果呈现为就像对卷积层特征图进行了有效压缩操作，把有用的信息按照相对位置进行提取，最终得到的特征图的大小肯定是变小了很多，但是有效保留了有用信息。

第一个下采样层 S1 向前到第二个卷积层 C2，该层由 8 个特征图组成，卷积滤波器大小设置为：36×26。C2 特征图与 S1 所有的特征图相连，实验中间结果如图 3-14 所示。

从下面的中间结果图我们可以看出，像上面的卷积处理效果一样，8 个特征图从不同的角度对上面的下采样层进行分析。不同的是，这里的卷积是对上面的下采样层进行处理，又因为下采样层是经过之前的不同角度的分析得到的中间结果，所以这里卷积的操作是在不同角度的特征图上进一步地抽象。

第二个卷积层继续向前到第二个下采样层，该层的下采样窗口大小设置为：3×3，实验中间结果如图 3-15 所示。

从下面的中间结果可以看出，八个下采样层特征图整体尺寸减小不少，但是特征图的内容和上面的八个卷积层特征图的内容很相似即进行了重要内容的蒸馏处理，保留了相对位置信息。

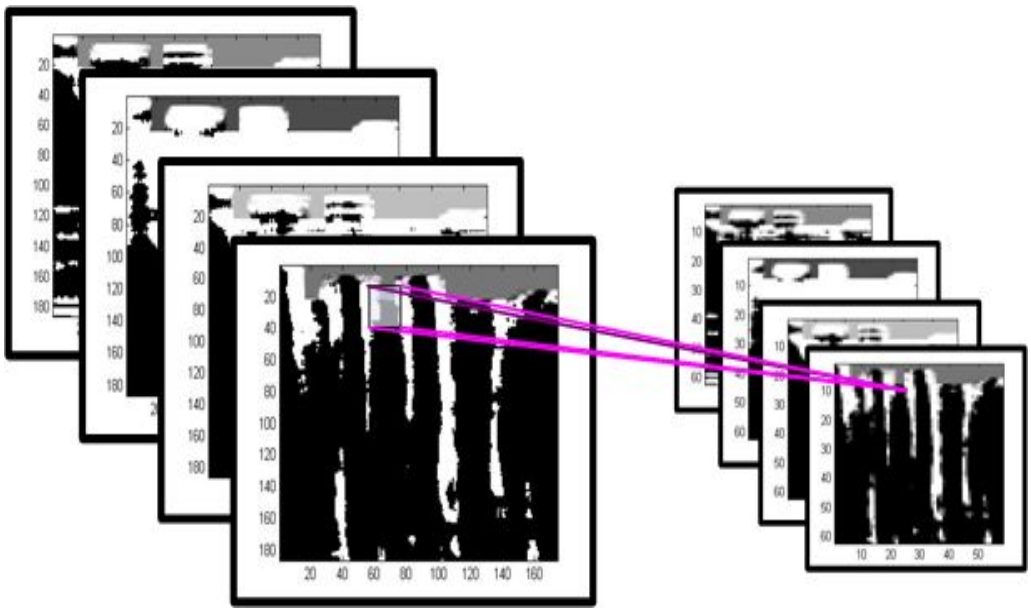

图 3-13 卷积层-下采样层

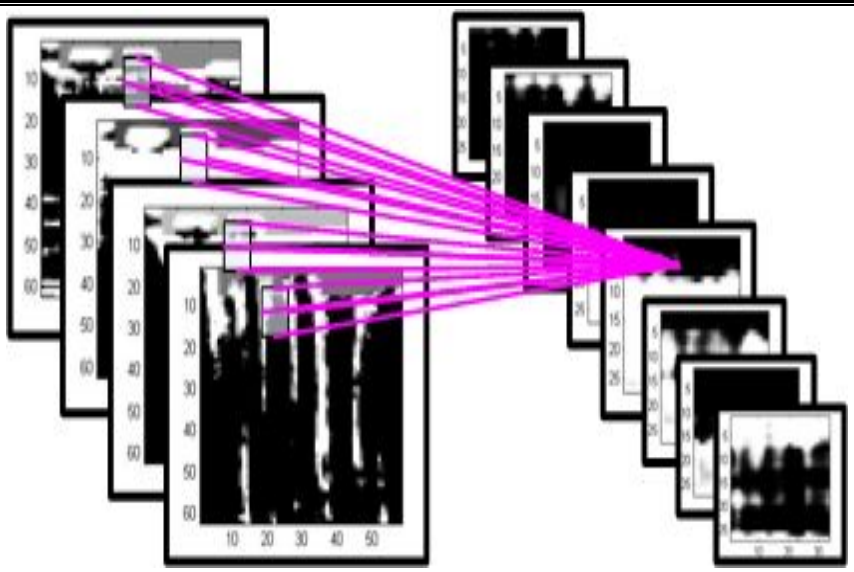

图 3-14 下采样层-卷积层

图 3-15 卷积层-下采样层

第二个下采样层继续向前为输出层，该层神经元的数量为场景类别数量。该层和输出层之间的连接方式像普通的 BP 网络一样为结点之间全连接。

至此，就完成了音频场景数据的特征图的计算，把第 2 个下采样层对应的特征图作为最终识别用的声学特征向量，它是大小为 9×11 的二维矩阵。通过调节 KNN 分类器的参数值，得到最终的识别结果如图 3-16 所示。实验系统的整体识别率为 0.55，基线系统的整体识别率为 0.52。可以看出，通过 CNN 提取出的长时结构性特征向量，相比较 MFCC 而言对音频场景识别来说是更有效的。

此外，如果以第 2 个下采样层为输入层连接一个 Softmax 分类层，进行音频场景的识别，得到的识别结果是：0.72，明显好于基线系统的识别结果，如图 3-17 所示。

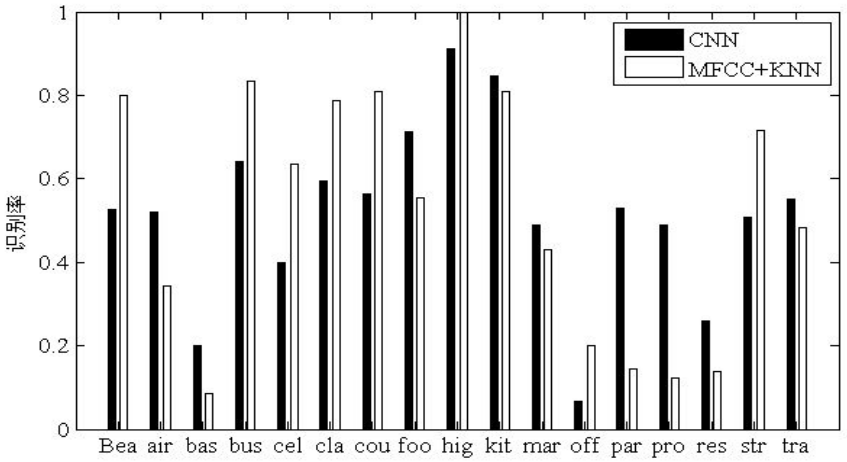

图 3-16 基于卷积神经网络的场景识别结果

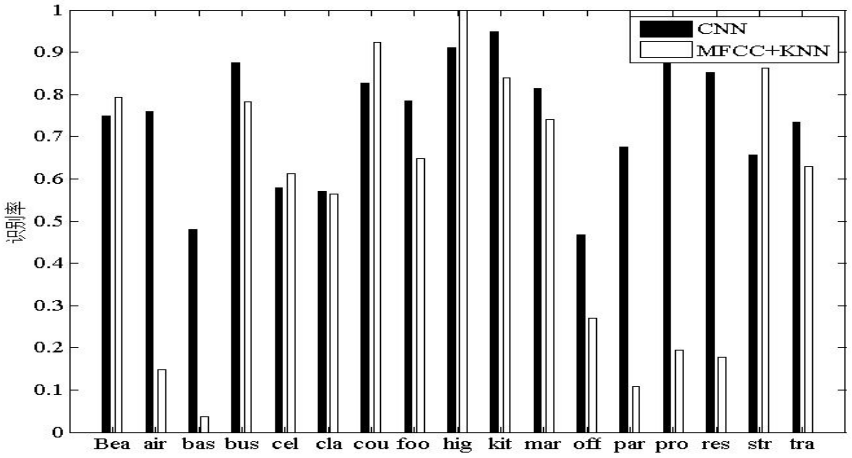

图 3-17 Softmax 分类器场景分类结果

3.6 本章小结

本章先是指出现有的音频特征用于音频场景识别存在的不足之处，然后提出了从基于频域的长时分析角度对音频场景数据进行特征提取的方法。由于神经网络在图像处理方面特别是对其结构性特征分析方面取得了很好的效果，所以我们应用卷积神经网络在语谱图上进行长时结构性特征分析。通过设计合理的网络结构和调整影响网络性能的多个参数，甚至是不断调整网络的局部结构，最终得到了好于基线系统的识别效果。

第 4 章 基于解卷积神经网络的音频特征提取及场景识别

4.1 引言

上一章我们首先分析了已有音频特征的不足，然后提出基于卷积神经网络的音频场景特征提取算法，按照算法提取的长时结构性特征在音频场景识别实验中取得了好于基线系统的成绩，验证了长时结构性特征的有效性。

深层神经网络模型之所以具备自学习能力，那是因为它对大量的样本数据进行分析，从中提炼出最具代表性的特征内容。所以，如果想要取得较好的效果，往往需要大量的训练样本。如果训练样本数量不够的话，不但有可能得不到很好的效果，更大可能是结果会很差。虽然上一章实验系统取得了好于基线系统的识别性能，但是训练样本数目不足一万，这对于实验系统所用到的网络模型来说是远远不够的。又因为 CNN 的训练需要带标签数据，现实情况往往是没有那么多的带标签的数据，因为类别标注是一件非常耗时间的工作，然而音频语料的采集工作相对来说较容易进行。所以，本章我们使用解卷积神经网络模型来进行音频场景的特征分析提取，因为它的整个训练过程无需数据的类别标签，这样就较好的解决了上一章训练数据不足的问题。

本章研究的内容是使用解卷积神经网络(DeConvolutional Neural Networks, DeCNN)对音频场景信号的语谱图进行特征分析，然后利用得到的特征向量进行音频场景识别。

4.2 基于解卷积神经网络的特征分析

DeCNN 和 CNN 对输入数据进行非线性变换的实现方式是相似的，即它们都是通过卷积操作和池化操作实现的；不同的是 DeCNN 的卷积运算的方向和 CNN 是相反的,而且还有些处理方式也是 DeCNN 独有的;这也让它具有了 CNN 不具有的能力，比如可以对原始输入数据进行重构。基于 DeCNN 特征分析的场景识别整体框架图如图 4-1 所示。

本小节主要是简单地介绍下 DeCNN 的理论知识。和上面一章相似，首先通过简单的拓扑结构图对 DeCNN 进行直观意义上的说明；然后通过 DeCNN

的计算方式，介绍下在给定训练好的模型的基础上，如何进行输入数据的特征提取；最后介绍下 DeCNN 的参数训练过程。

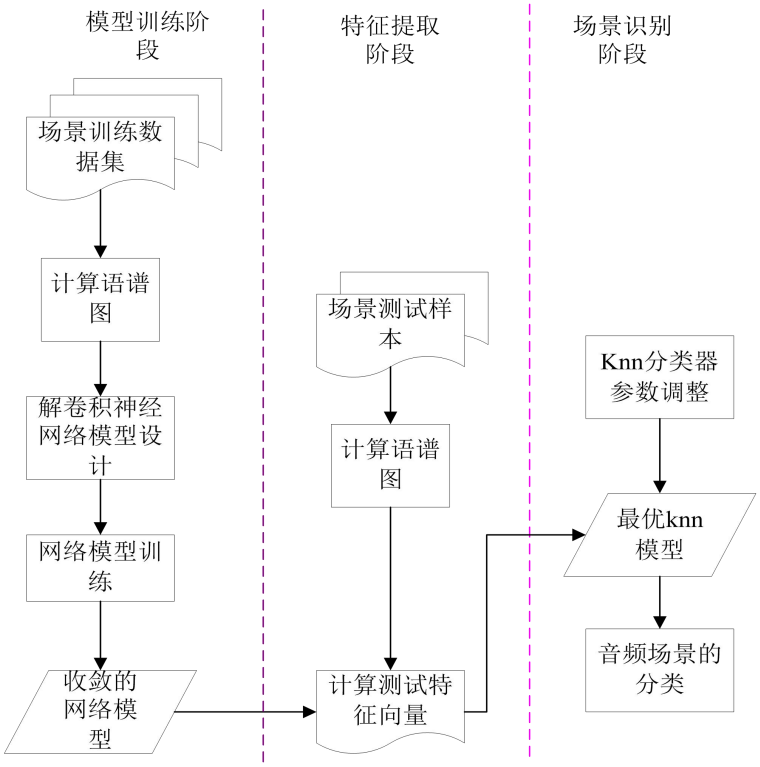

图 4-1 基于解卷积神经网络特征提取的场景识别整体框架图

4.2.1 解卷积神经网络的拓扑结构

如图 4-2 所示，为一个简单的两层 DeCNN 的拓扑结构图。从拓扑结构图中我们可以看出 DeCNN 是通过几个卷积稀疏编码层即解卷积层和最大池化层的间隔设置，把输入数据进行非线性分解。每一个解卷积层都由几个特征图组成，这一点和 CNN 相似；不同的是每个解卷积层都尝试在稀疏性约束的前提下对最小化输入数据的重构误差这一目标进行优化。

4.2.2 解卷积神经网络的计算方式

首先说明一下解卷积神经网络在对目标函数进行优化过程中用到的代价函数 $C(y)$，如式(4-1)所示。其中可以看出代价函数由两部分组成：似然项和正则化项[49]。其中似然项设置的目的是为了让重构结果尽可能向输入数据靠近；正则化项则是对整个模型复杂度的控制，以防止过拟合现象的发生。最后，用一个参数对两项的相对权重进行调节。

$$C_l(y) = \frac{\lambda_l}{2}\left\|\hat{y_l} - y\right\|_2^2 + \sum_{k=1}^{K_l}\left|z_{k,l}\right|_1 \tag{4-1}$$

上式中：λ_l为两项的调节参数，$\hat{y_l}$为第l层的重构结果，y为输入数据，$z_{k,l}$为第l层中的第k个特征图，K_l表示第l层特征图的数量。

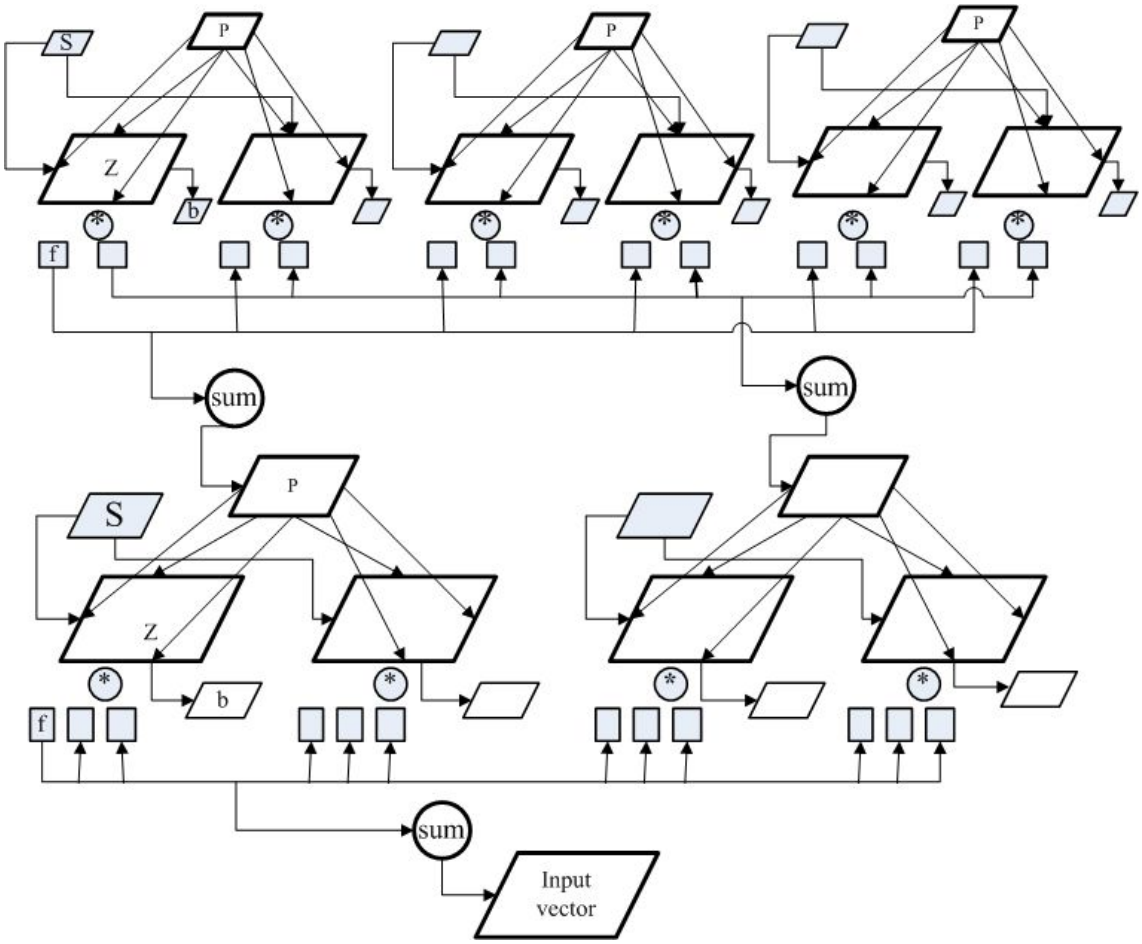

图 4-2 解卷积神经网络拓扑结构示意图

从上图我们可以看出 DeCNN 和按照常规思路进行重构的神经网络不同的是，DeCNN 的所有解卷积层都是对输入数据进行重构，即都是以最小化重构输入数据为优化目标，而不是致力于对下层的输入向量进行重构。

我们本实验系统的目的是使用 DeCNN 对音频场景数据的语谱图进行深层特征提取，所以进行 DeCNN 模型训练的最终目标就是对特征图的计算。下面是在给定音频场景的语谱图和卷积滤波器初始值的前提下，对特征图的求解以及池化位置变量的更新过程。

如图 4-3 所示为 DeCNN 的整个计算过程，我们可以看出计算过程的主线就是以特征图为基础对原始输入内容进行重构计算，第l层对输入数据进行重

构的计算过程如下：使用卷积滤波器根据该层以下所有层次中的池化位置变量和该层的特征图 z_l 按照公式(4-2)进行重构操作，最终得到的重构结果为 $y_l^{\wedge}$。

$$y_l^{\wedge} = F_1 U_{s_1} F_2 U_{s_2} \ldots F_l z_l = R_l z_l \tag{4-2}$$

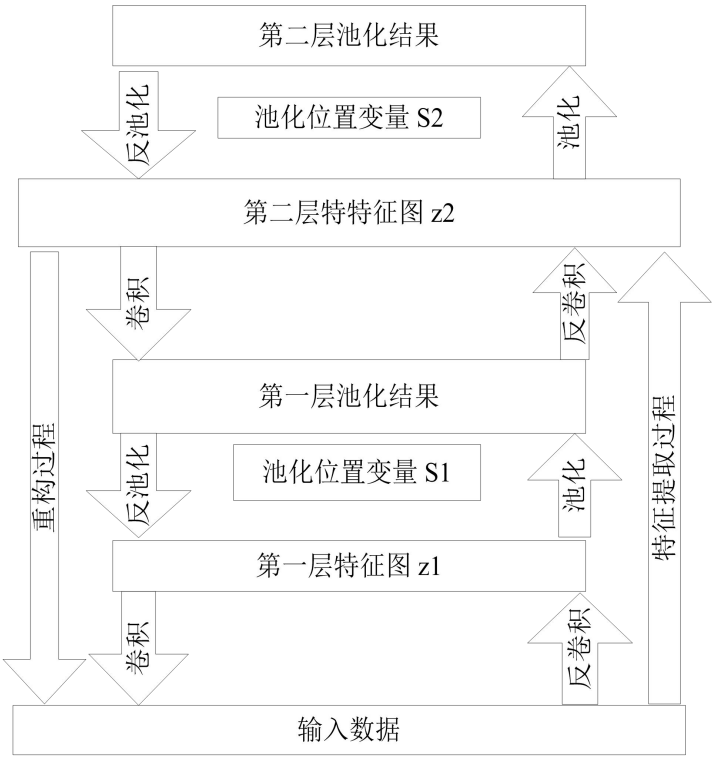

图 4-3 DeCNN 计算过程示意图

这里对上式进行简单的说明。它不是严格意义上的数学表达式，而是一个流程表达式。其中 F_l 表示对第 l 层特征图进行卷积操作，U_{s_l} 表示根据第 l 层对应的池化位置变量 s_l 进行反池化操作。这里的卷积操作和 CNN 里的操作方式相同，就不再介绍。反池化操作则是根据池化位置变量把池化结果向特征图还原的过程，后面会对其详细介绍。从上式可以看出，从第 l 层开始经过卷积操作和反池化操作一直到第一层，最终得到第 l 层对输入数据的重构结果。$R_l z_l$ 是根据第 l 层特征图 z_l 对输入数据进行重构操作的整个流程的简写形式。

有了上面的重构结果 $y_l^{\wedge}$，可以计算出重构误差 $y_l^{\wedge} - y$，进而可以通过重构误差的反向传播对特征图进行更新，最终得到我们所求的特征图。可以看出，这里的反向传播和 CNN 中的反向传播的方向正好相反，即它是从输入数据开始向特征图进行传播的。

整个计算过程我们按照迭代阈值收缩模式[50]进行，整个迭代过程由：梯度的计算、特征值收缩和池化操作三个主要操作步骤组成，下面对这三个主要操作进行简单的说明。

（1）梯度计算：

梯度计算这一步主要是进行代价函数$C(y)$关于特征图z_l的梯度的求解。梯度计算涉及到的操作和上面的重构计算涉及到的操作是相反的，它是通过反卷积操作和池化操作进行的，计算公式如(4-3)所示。

$$R_l^T = F_l^T P_{s_{l-1}} F_{l-1}^T P_{s_{l-2}} ... P_{s_1} F_1^T \tag{4-3}$$

上式也是一个流程表达式，它和式(4-2)是完全相反的两个流程。其中F_l^T表示对第l层的特征图进行反卷积操作，和 CNN 中反向传播中的反卷积计算方式相同，这里就不再介绍；P_{s_l}表示对第l层特征图进行池化操作，并记录池化位置变量s_l，其中的池化操作和 CNN 中的池化操作相同。可以看出，从第一层输数据开始经过反卷积和池化操作，一直到第l层就得到了第l层的特征图。所以，第l层的重构误差关于第l层特征图的梯度计算，就是通过把重构误差$(R_l \mathrm{z}_l - \mathrm{y})$从第一层通过上面的整个过程传递到第$l$层进行计算的，最终得到的梯度计算结果如公式(4-4)所示，其中R_l^T表示对第l层的特征图的整个计算过程的简写。

$$g_l = R_l^T (R_l \mathrm{z}_l - \mathrm{y}) \tag{4-4}$$

有了上面的梯度结果，特征图的更新就是按照公式如(4-5)所示进行调整。

$$z_l = z_l - \lambda_l \beta_l g_l \tag{4-5}$$

上式中，β_l为学习速率。

从式子(4-2)我们可以看出，第l层对原始输入数据进行重构的计算过程仅和本层的特征图z_l有关，和该层以下的层次没有关系。事实上，重构操作R_l的执行和下面所有层次的池化位置变量：$(\mathrm{s}_{l-1}, \mathrm{s}_{l-2} ... \mathrm{s}_1)$都是相关的，因为执行反池化操作需要依赖这些变量的值。然而这些位置变量的更新又和下面层次的特征图有关，计算公式如(4-6)所示，所以第l层重构操作和该层以下的所有层次的特征图都是有关系的。

$$[p_l, \mathrm{s}_l] = P(\mathrm{z}_l) \tag{4-6}$$

上式中$P()$是池化操作，p_l为经过池化操作得到的特征图，s_l为记录下的池化位置变量。

（2）收缩操作：

完成上面的梯度计算步骤之后紧接着就是特征图收缩操作，该操作是通过把特征图中较小的元素值置零实现的。该操作的主要目的就是增加特征图的稀疏性，计算公式如(4-7)所示。

$$z_l = max(|z_l| - \beta_l, 0) \mathrm{sign}(z_l) \tag{4-7}$$

上式中，sign 为符号函数，*max* 表示取最大值操作。

(3) 池化反池化操作：

该操作是发生在特征图和池化结果之间的，通过对特征图的池化操作我们得到池化结果，对池化结果进行反池化操作可以得到特征图。其目的主要是为了更新池化位置变量，进而得到一个对应该层次的最优池化位置变量，详细过程下面介绍。

4.2.3 解卷积神经网络的学习算法

从上面的计算过程我们可以看出，DeCNN 训练过程中需要学习的参数有卷积滤波器 f，同时训练过程中需要计算的变量有特征图 z 和池化位置变量 s。

通过求解损失函数关于 f 的导数并令其等零，可以得到 $y_l^{i\wedge}$ 和 y_l^i 相等，进而可以得到如下线性系统，通过求解该线性系统可以对 f 进行更新。

$$\sum_{i=1}^{N}(z_l^{iT} P_{s_{l-1}}^i R_{l-1}^{iT}) y_l^{i\wedge} = \sum_{i=1}^{N}(z_l^{iT} P_{s_{l-1}}^i R_{l-1}^{iT}) y_l^i$$

给定输入数据 y 和随机初始化的卷积核 f，DeCNN 的学习过程如算法 4-1 所示。

算法 4-1 DeCNN 学习算法

输入：训练集合 S、网络层次 L、迭代次数 E、正则化系数 λ、收缩参数 β、ISTA 步数 T

输出：卷积核 f、特征图 z、池化特征位置 s

for l=1: L

按照两个变量都服从高斯分布的原则，随机初始化卷积核 f_l，特征图 z_l。

for epoch = 1 : E

for i =1 : N

for t = 1 : T

1）重构结果计算：$y_l^{\wedge} = R_l z_l^i$；

2）重构误差计算：$e = y_l^i - y^i$；

3）损失函数关于特征图的梯度计算：$g_l = R_l^T e$；

4）特征图更新：$z_l^i = z_l^i - \lambda_l \beta_l g_l$；

5）特征图收缩操作：$z_l^i = \max(|z_l^i| - \beta_l, 0)\text{sign}(z_l^i)$；

6）对特征图 z_l^i 进行 pool 操作，按照下式更新特征位置 s_l^i；

$[p_l^i, s_l^i] = P(z_l^i)$

7）使用特征位置 s_l^i，对特征图 p_l^i 进行 unpool 操作计算；

$z_l^i = U_{s_i} p_l^i$

End for

End for

使用共轭梯度法通过求解上面的线性系统来更新卷积滤波器： f；

End for

End for

最终输出：收敛模型对应的卷积滤波器 f，特征图 z 以及池化位置变量 s。

4.3 解卷积神经网络用于音频场景的特征分析及识别算法

通过对解卷积神经网络的训练，我们得到收敛的模型，进而利用它提取音频场景特征，最终借助 KNN 分类器进行最后的音频场景识别。本小节我们首先进行 DeCNN 在音频特征提取方面的适用性分析，然后给出基于 DeCNN 的特征提取算法，最后利用提取到的特征向量进行音频场景识别。

4.3.1 解卷积神经网络的适用性分析

上一章中我们介绍的 CNN 以它独有的卷积操作和下采样操作在特征提取方面表现出了很好的性能，而 DeCNN 保留了 CNN 的卷积和下采样操作，所以卷积操作和下采样操作对特征分析带来的帮助在这就不重复介绍。但是 DeCNN 不同于 CNN 的地方是它每层的特征图都是对原始数据进行重构，这就相当于 DeCNN 的训练学习过程是对数据原始内容进行编码解码操作，即正向传播对应为解码操作，反向传播对应为编码操作。

之所以具备对输入数据进行编解码的功能是因为解卷积神经网络的下面两个独特的操作：

(1) 解卷积操作：从上面的拓扑图可以看出网络中的任意一层对输入数据的重构结果是由特征图和卷积滤波器进行解卷积操作之后再求和得到的，计算公式如(4-8)所示。

$$\hat{y_1} = \sum_{k=1}^{K_1} z_{k,l} * f_{k,l} \tag{4-8}$$

上式中的*为二维卷积操作，其中卷积核 f 对所有的输入数据都是共有的，

但是特征图*z*是个隐变量,它和输入的语谱图是一一对应的，是网络训练过程中待学习的参数。该操作之所以被称为解卷积，是因为它和 CNN 中类似的操作有所不同，它是对特征图进行卷积进而得到重构结果。我们可以看出特征图的数目是大于 1 的，不同的特征图从不同角度对输入数据进行分析，所以多个特征图保证了模型学到的特征的完备性，同时其中的正则化又限制了模型的复杂度，避免了过拟合现象的发生。对音频场景信号的语谱图进行解卷积操作，就相当于每层的特征图都是对语谱图的一个编码操作。和以往的编码操作不同的是 DeCNN 的编码是借助池化位置变量来完成的，而该变量和输入语谱图是一一对应的。进行解卷积操作的对象是特征图中特定的神经元，它对应了输入语谱图中的一小片区域；而这种对应关系和语谱图对特征单元值的生成所作出的贡献相一致的。又因为从语谱图向特征图映射是通过卷积运算和下采样运算来完成的，所以这种通过解卷积操作来完成的编码结果就会保留下长时结构性特征。

(2) 池化、反池化操作：从上面的拓扑结构图我们可以看出 DeCNN 的池化为三维池化，对应 CNN 中的这一操作是二维池化。这样操作的好处就是可以在比当前层次更大的粒度上来进行结构特征的分析，进而得到更具结构性的特征。此外，我们在池化操作过程中不仅得到池化操作的结果，而且还记录了特征神经单元在特征图中的位置。其中包括特征图的下标和对应特征图内部的坐标位置，我们称它为池化位置变量(switches)。借助这个池化位置变量，反池化操作把特征图中的元素值设置为池化结果图中对应位置的值，同时把特征图中其它元素的值置零，过程如图 4-4 所示。

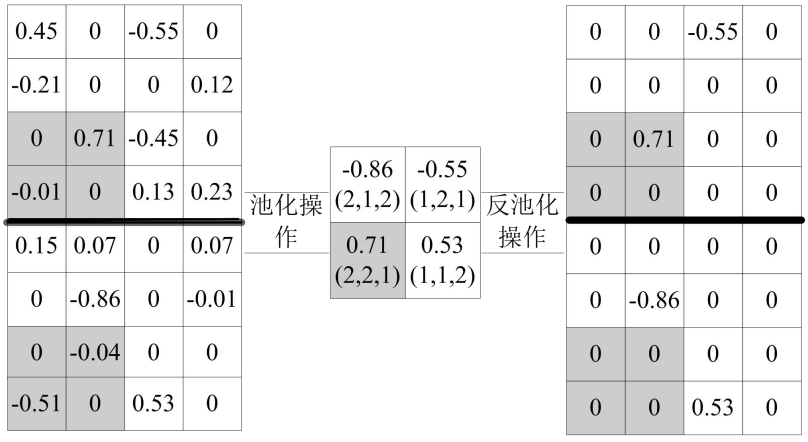

图 4-4 池化反池化操作示意图

很明显，经过池化操作，原始输入数据对应的语谱图中的有效内容被保留了下来，相反那些不相关的内容就被丢弃。这会使语谱图中有效成分的绝对位置被模糊，突出了结构性特征的相对位置的重要性。反池化操作不仅保留了语

谱图中有效成分的结构，而且使重构得到的结果更加稀疏，对抽取输入数据语谱图的结构性特征更加有利。

4.3.2 解卷积神经网络用于音频场景特征提取及识别算法

上面我们已经分析了 DeCNN 在特征提取方面独特的能力，接下来是利用 DeCNN 对音频场景数据进行深层特征分析以及最后的场景识别，算法如下：

(1) 音频场景数据预处理，得到对应的神经网络输入变量。

(2) 设计 DeCNN 拓扑结构，实验系统所用到的 DeCNN 为 9 层的网络模型。

(3) 初始化卷积滤波器为较小的随机值，然后根据上面的学习算法进行迭代更新。

(4) 得到收敛的 DeCNN 模型后，取每层的特征图作为音频场景信号的特征向量。

(5) 利用上面得到的特征向量训练分类器，然后进行音频场景的分类。

4.4 实验结果及分析

4.4.1 实验数据预处理

本章实验所用的音频场景语料和前面的相同，语料含有 17 个场景类别共计 8298 个样本。与 CNN 进行训练时数据的使用有所不同的是，对 DeCNN 的训练我们使用所有的数据来进行。经过时频分析转换之后得到的语谱图的大小为：221×198，我们把它规整化为 221×221。

4.4.2 实验基本网络结构

在确定网络模型结构过程中，主要对影响场景识别性能较大的几个参数：网络层次的数目和卷积滤波器尺寸进行了调整，调整内容和最终的识别性能关系如图 4-5 所示。

最终，实验中设计的网络结构共 9 层，卷积核大小设置为：7×7，三维池化中下采样的窗口大小为 3×3×2，即从特征图中 3×3 共计 9 个神经元中取出最大值，然后在三个相邻的特征图之间再进行一次取最大值操作，相当于在 27 个神经元中取最大值作为最终的特征元素。这不仅仅是简单的神经元数量上的增加，更重要的是特征图数量，即特征分析角度上的变化，整体网络结构图如图 4-6 所示。

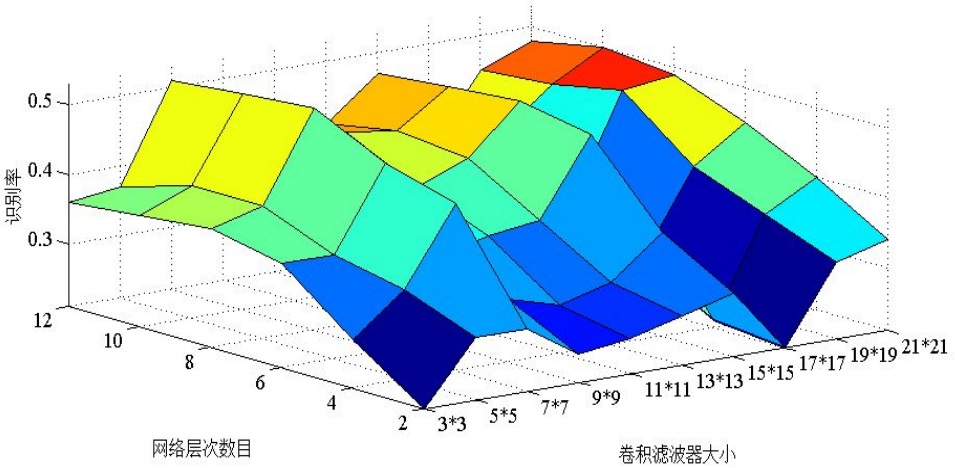

图 4-5 网络参数和识别性能关系示意图

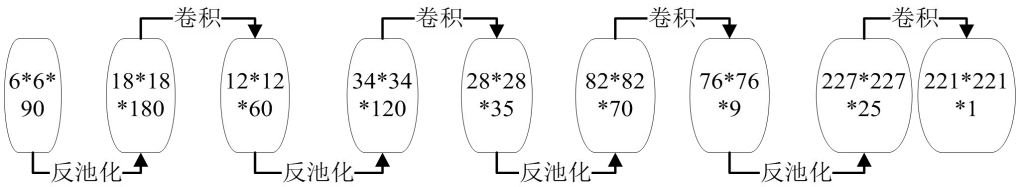

图 4-6 实验网络结构示意图

4.4.3 实验参数调整及结果分析

解卷积神经网络经过训练达到收敛状态之后，我们就得到了进行最终场景识别所用到的特征图，我们对特征图做一些小的处理和原始的特征图进行最终的识别对比实验。实验对第二层特征图进行调整，按照同时只改变一个变量的原则，实验分两组进行，即池化特征图一组，反池化特征图为另一组。

调整方案大致如下：作为对比实验我们选择原始特征图作为最终分类用到的输入特征向量；调整策略是把特征图中一定比例的元素固定不变，其它元素都置零，该部分元素我们选择的原则是绝对值较大的那一部分，元素比例按照：1/2、1/4、1/8 三种调整方案进行试验，第一、二组的实验结果如图 4-7 所示。

从下面的实验结果可以看出，无论是小组一还是小组二如果取全部元素作为最终分类实验的特征向量，然后进行场景识别的结果都是最好的。当元素的抽取比例逐渐下降时，可以看出整体的识别率逐渐下降。这说明元素比例越大，对应特征的完备性越强。当抽取的元素比例为 1/2 时，小组一中整体识别效果几乎和全部元素作为特征值时相当，小组二中的整体识别率略有下降但是相差不大。然而这样处理的话，特征向量整体的稀疏性增加了很多，相应的鲁棒性

也会增加。所以，我们可以在元素比例为 1/2 和 1 之间进行取舍作为我们最后的特征向量。

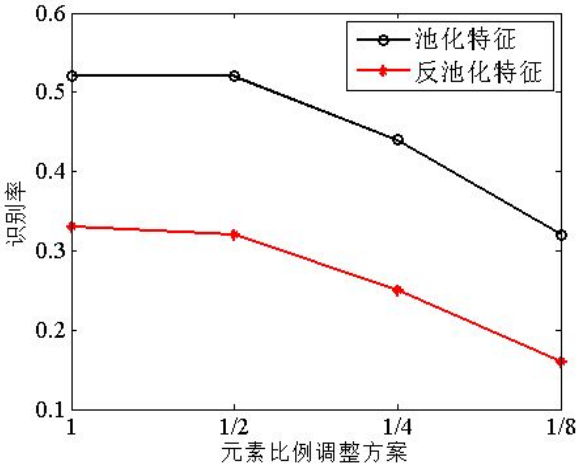

图 4-7 元素比例调整结果示意图

本章实验系统的整体类别识别率和基线系统对比结果图如图 4-8 所示。从下面的结果图中我们可以看到，实验系统和基线系统在场景类别识别率上各有自己的高点，同时也有自己的低谷，实验系统整体识别率比基线系统识别率略微提高。

DeCNN 的整个训练过程没有任何输入数据的类别标签参与，通过对输入数据的重构操作完全自适应地对参数进行调整，这种无监督的训练方式本身就需要大量的训练数据进行训练才会取得较好的成绩。虽然实验系统的整体识别性能没有较大的提升，但是实验系统为解决现实情况下数据标签不足的问题提供了解决方法。如果增加训练数据样本数量的话，理论上实验系统会取得更好的识别性能。

4.5 本章小结

本章先是简单的阐述了卷积神经网络进行训练过程中存在的不足之处，然后提出了使用解卷积神经网络模型进行音频场景识别的方法。整体网络模型和卷积神经网络相似，也是卷积层和池化层交叠设置。但是计算方式是从上向下进行卷积，卷积方向和池化方向相反，同时设置了自己独特的反池化操作。虽然整体识别效果较卷积神经网络没有较大提升，但是却解决了卷积神经网络训练语料不足的问题。

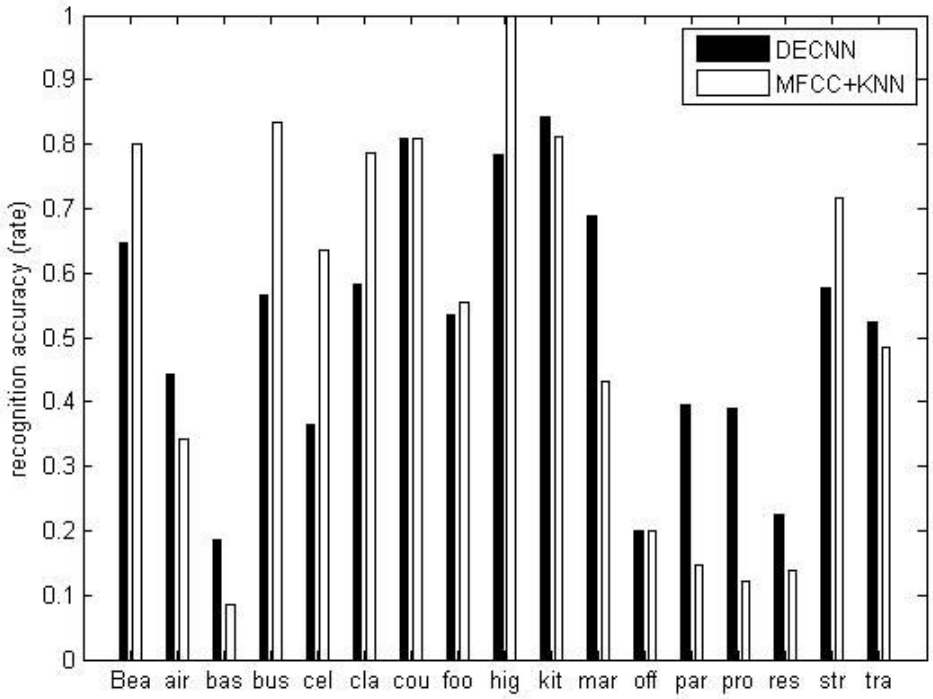

图 4-8 基于解卷积神经网络的音频场景识别结果

结　论

语音信号处理作为人工智能重要研究方向之一，为人机交互等智能行为的实现有着重要的意义。而其中的音频场景识别对于智能机器感知外界环境的场景类别，进一步做出智能化动作有着十分重要的作用。已有的场景识别方法都是按照先特征提取后分类器分类的研究框架进行的。但是，音频场景信号和普通的音乐或者说话声等有所不同，它所含有的关键声学事件时长不一且平均时长较长，所以按照已有的音频特征分析的方式进行场景的特征分析会影响其识别性能。在这样的研究背景下，我们提出使用在结构性特征分析取得很好成绩的深层神经网络来对音频场景进行结构性特征分析，最终提高了场景识别性能。具体工作内容如下：

首先我们研究了基于卷积神经网络的音频场景的特征分析。其中我们分析了常规时域、频域或者倒谱域等音频特征在音频场景识别任务中存在的不足，又分析了长时统计特性因其结构性特征的丢失而存在的问题，最终我们对基于频域的长时分析数据进行场景特征进行分析。卷积神经网络对自然图像的结构刻画描述能力在很多方面都得到了验证，所以我们利用它结构性特征分析提取的能力在语谱图上进行长时结构性特征分析。通过最终的场景识别性能对比，说明这种长时特征的方法是很有效的。

然后我们基于上面的研究方案存在的不足之处，提出了改进方案。上面我们研究的卷积神经网络在性能上很突出，但是网络的训练需要大量的有标签数据，现实情况是场景语料容易采集，但是对其标注却很费时费力。所以，我们提出了使用解卷积神经网络进行场景数据特征提取的方法。解卷积神经网络模型保留了卷积和下采样操作，但是参数更新过程无需训练数据的类别标签，即它是通过对输入数据的重构操作来调节参数的。

通过实验，我们可以看到卷积神经网络在场景识别方面的性能有了很大的提高，解卷积神经网络虽然没有较大地提高识别性能但是却解决了训练数据不足的问题。两者都是很好的研究方向，但是两者都还有很多研究工作要做，下面是对其分析：

首先是神经网络训练过程中参数的调整。训练过程涉及到很多参数的调整，比如学习速率、权重衰减项系数、动量项等，此外还有网络结构方面的调整。其次是对语谱图中内含关键声学事件的结构性分析，以及对应的卷积滤波器尺寸的分析和不同层次间卷积滤波器的尺寸调整等都有进一步实验的空间。

参考文献

[1] Chu S, Narayanan S, C-C J K. Environmental Sound Recognition Using mp based Feature[C]. ICASSP，2008：1-14.

[2] Agostini G, Longari M, Pollastri E. Music instrument time bress classification with spectral feature[J]. EURASIP，2003(1)：5-14.

[3] Wei T C, Wen H C, Ja L W, Jane Y H. A study of semantic context detection by using SVM and GMM Approaches[C]. IEEE international conference，2004：1951-1954.

[4] Khin M C, Zin K L. Audio Based Action Scene Classification Using HMM SVM Algorithm[J]. IJARCET，2013.

[5] Heittola T, Mesaros A, Eronen A, Virtanen T. Audio Context Recognition using Audio Event Histogram[C]. In 18th European Signal Processing Conference：1272-1276.

[6] Heittola T, Mesaros A, Virtanen T, Eronen A. Sound Event Detection in Multi Source Envronments using Source Separation[C]. CHIME，2011：36-40.

[7] Mesaros A, Heittola H, Klapuri A. Latent Semantic Analysis in Sound Event Detection[C]. European Signal Processing Coference，2011.

[8] Pelto V, Tuomi J, Klapuri A. Computational Auditory Scene Recognition[C]. ICASSP，2002.

[9] Gueo E G, Li Sheng Z. Content based Audio Classification and Retrieval by SVM[J]. IEEE Trans. Neural Network，2003：209-215.

[10] A J, McCowan I, Bourland H. Speech Segmentation Using Entropy and Dynamic Features in a HMM Classification Framework[J]. Speech Common，2003：351-363.

[11] Hwang K, Lee S Y. Environmental Audio Scene and Activity Recognition through Mobile based Crowd sourcing[J]. IEEE Transaction on Consumer Electronics，2012，58(2)：700-705.

[12] Srinivasan S, Jin Z, Shao Y. A Computational Auditory Scene Analysis System for Speech Segregation and Robust Speech Recognition[J]. Computer Speech and language，2010，24(1)：77-93.

[13] Akinori I, Akihito A, Massashi I, Shozo M. Detection of Abnormal Sound Using Multi-stage GMM for Surveillance Microphone[J]. IAS， 2009：

733-737.

[14] Rosenblatt F. The Perceptron：A Probabilistic Model for Information Storage and Organization in the Brain[J]. Cornell Aeronautical Laboratory. Psychological Review，1958，65(6)：386-408.

[15] Minsky ML, Papert SA. Perceptrons[M]. Cambridge：MIT Press，1969.

[16] Cortes C, Vapnik V. Support Vector Networks[J]. Machine Learning，1995，20.

[17] Boser B E, Guyon I M, vapnik V N. A Training Algorithm for Optimal Margin Classifiers[C]. Proc of the 5th Annual ACM workshop on COLT. Pittsburgh，PA，1992，144-152.

[18] Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V. Support Vector Regression Machines[C]. NIPS，1996：155-161.

[19] 张学工译. 统计学习理论的本质[M]. 北京： 清华大学出版社，2000.

[20] Freund Y, Schapire Re. A Short Introduction to Boosting[J]. Journal of Japanese Society for Artificial Intelligence，1999，14(5)：771-780.

[21] LeCun Y, Boser B, Denker J S, D Henderson, Howard R E, Hubbard W, Jackel L D. Handwriting Digit Recognition with a Back Propagation Network[C]. NIPS，89.

[22] Hinton G. E, Osindero S, Teh Y W. A Fast Learning Algorithm for Deep Belief Networks[J]. Neural Computation，2006，18：1527-1554.

[23] Hinton G E, Salakhutdinov R. Reducing the Dimensionality of Data with Neural Networks[J]. Science，2006，313：504-507.

[24] Bengio Y, L P, P P, L H. Greedy Layer Wise Training of Deep Belief Networks[C].NIPS，2007.

[25] Abdel R M, George E, Hinton G E. Acoustic Modeling using Deep Belief Networks[C]. IEEE transaction on audio speech and language processing，2010.

[26] Hinton G E, Sejinowski T J ed. Learning and Relearning in Boltzmann Machines [J]. Microstructure of Cognition，1986，1：282-317.

[27] Salakhutdinov R, Hinton G E. Deep Boltzmann machines[C].International Conference on Artificial Intelligence and Statistics，2009：448-455.

[28] Sriva N, S R. Multimodal Learning with Deep Boltzmann Machines[J]. Machine Learning，2014.

[29] George D, Dong Y, Li D, Alex A. Context-Dependent Pre-trained Deep Neural networks for Large Vocabulary Speech Recognition[J]. IEEE

Transactions on Audio Speech and Language Processing，2012，20：30-42.

[30] Hinton G, L D, Dong Y. Deep Neural Networks for Acoustic Modeling in Speech Recognition[J]. IEEE Signal Processing，2012，29：82-97.

[31] George D, Dong Y, Li D, Alex A. Large Vocabulary Continuous Speech Recognition With Context-dependent DBN-HMMS[C]. ICASSP，2011.

[32] krizhevsky A, Hinton G E. Imagenet Classification with Deep Convolution Neural Networks [C]. NIPS，2012.

[33] Graves A, Mohamed A R, Hinton G. Speech Recognition with Deep Recurrent Neural Networks[C]. ICASSP，2013.

[34] Mesnil G, Xi D H, L D, Yoshua B. Investigation of Recurrent Neural Network Architectures and Learning Methods for Spoken Language Understanding[C]. Interspeech，2013.

[35] Hannun A, Case C, Casper J, Andrew Y Ng. DeepSpeech: Scaling up end-to-end Speech Recognition[C]. arxiv，2014.

[36] Yaniv T, Ming Y, Marc' Aurelio Ranzato, Wolf L. DeepFace: Closing the Gap to Human Level Performance in Face Vertificaiton[C]. CVPR，2014.

[37] Dong Y, Shi W, Li D. Sequential Labeling Using Deep Structured Condition Random Fields [J]. IEEE Journal of Selected Topics in Signal Processing，2010，4(6)：965-973.

[38] Honglak Lee, Largman Y, Pham P, Andraw Y NG. Unsupervised Feature Learning For Audio Classification Using Convolutional Deep Belief Networks[C]. NIPS，2009：1096-1104.

[39] Honglak Lee, Roger G, Rajesh R, Andrew Y Ng. Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations[C]. ICML，2009.

[40] 韩纪庆，张磊，郑铁然. 语音信号处理[M]. 北京：清华大学出版社，2013.

[41] 李航. 统计学习方法[M]. 北京：清华大学出版社，2012.

[42] Michel M D, Elena D. Encyclopedia of Distances[J]. Springer，2009.

[43] Gonde G, Hui W, Bell D, Bi Y, Greer K. KNN Model Based Approach in Classification[C]. OTM Confederated international conference, CoopIS, DOA, ODBASE，2003，2888：986-996.

[44] Zhang M, Zhou Z. ML-KNN: A Lazy Learning Approach to Muti-Label Learning[J]. Pattern Recognition，2007(7)，40(7)：2038-2048.

[45] Hall P, Samworth B J. Choice of Neighbor Order In Nearest Neighbor Classification[J]. The Annals of Statistics，2008(10)，36(5)：2135-2152.

[46] Boyd S, Vandenberghe L. Convex Optimization[M]. 北京：清华大学出版社，2013.

[47] 陈宝林. 最优化理论与算法[M]. 北京：清华大学出版社，2011.

[48] Richard O D, Peter E H, David G S. Pattern Classification[M]. 北京：机械工业出版社，2012.

[49] Mattew D Z, Graham W, Taylor, Rob F. Adaptive Deconvolutinal Networks for Mid and High Level Feature Learning[C]. ICCV，2011.

[50] Beck A, Teboulle M. A Fast Iterative Shrinkage threshold Algorithm for Linear Inverse Problems[J]. SIAM，2009，2(1)：183-202.

哈尔滨工业大学学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《基于深层神经网络的音频特征提取及场景识别研究》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：  日期： 2015 年 6 月 30 日

学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其它复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其它成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：  日期： 2015 年 6 月 30 日

导师签名：  日期： 2015 年 6 月 30 日

致　谢

两年的研究生学习生活转眼就这么过去了，这两年里我有付出有收获。期间我学到了很多，不仅有学术方面的科学知识而且还有处理问题方面的方法技巧。感谢这两年里我遇到的每一个人，因为他们我的青春才更加精彩。

首先衷心地感谢我的导师郑铁然副教授，研究生生涯虽然只有短短的两年时间，但是郑老师对工作的专注精神和奉献精神对我的影响足以让我终生受益。感谢韩纪庆老师，郑贵滨老师，马心慧老师。韩老师以他严谨的治学态度、诲人不倦的教育方式在我两年的研究生学习生活中起到了非常重要的作用，郑老师的和蔼和马老师的细心、热情都给与我极大的精神鼓舞。

感谢实验室的所有同学们，尤其是师兄们在学术上给我的悉心指导让我在自己的学术积累过程中少走了很多弯路，为我的课题研究的顺利进行提供了很大的帮助。和同学们在一起度过的这两年快乐的时光将会是我终生难忘的美好回忆。

最后我要感谢我的父母，正是有了他们的无私支持我才有勇气一步一步向前走，感谢所有帮助过我的人，谢谢你们！