信息科学与技术学院

2016 级硕士研究生毕业论文

开题报告

题　　目：基于卷积神经网络的音频场景分类方法研究

专　　业：信息与通信工程

导　　师：李灿平

姓　　名：孙凌山

学　　号：2016020673

2018 年　6 月　29 日

目录

基于卷积神经网络的音频场景分类方法研究

1. 研究的目的及意义

通过分析声音使设备能够理解其环境是音频场景分类（Acoustic Scene Classification，下文简称ASC）研究的主要目标，该目标涉及计算听觉场景分析（CASA）。机器听音系统为人类听觉系统执行类似的处理任务，且通过机器学习，机器人技术和人工智能等领域试该主题的研究更进一步。

ASC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（IPA）也是一个受到ASC推动的领域。 IPA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。 IPA服务，如Google的Google Now、微软的Cortana、Apple的Siri以及亚马逊的Alexa广泛使用音频输入。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。

此前音频场景分类仍然基于将通用分类器（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（Convolutional Neural Network, CNN）是深度学习技术中的一种，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。卷积神经网络是一种前馈神经网络，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。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方面能够给出更好的结果。这一模型也可以使用反向传播算法进行训练。相比较其他深度、前馈神经网络，卷积神经网络需要的参数更少，使之成为一种颇具吸引力的深度学习结构。

2015 年,Piczak 等人（Piczak，2015）对深度学习中的卷积神经网络是否可有效的应用于音频场景分类这一问题进行了探讨。为此他们依照此前将卷积神经网络成功用于图像分类的经验运用于音频场景分类上。实验结果表明，使用卷积神经网络进行音频场景分类是一个切实可行的办法。卷积神经网络模型胜过基于手动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。且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。

2. 音频分类方法的研究现状

2.1. 一般音频分类方法的研究现状

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（Sawhney，Maes，1997）。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，1998） 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。

2.2. 基于卷积神经网络音频分类方法的研究现状

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，它在评估集中达到了83.3％，在DCASE 2016挑战任务中排名第35。

瓦伦蒂等人（Valenti等，2016）使用基于log-mel谱图的CNN，系统选择的特征表示是log-mel谱图。 为了计算log-mel谱图,他们在40ms的音频窗口上应用一个短时傅里叶变换（STFT），并重叠50％和Hamming窗口。 然后计算每个箱的绝对值并应用一个60段的梅尔比例滤波器组。 最后，计算mel能量的对数转换。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仓，两者都是在每次折叠的整个训练集上计算的。 然后，将归一化的光谱图分成更短的光谱图，在后面调用序列。与用于STFT的帧不同，他们选择序列不重叠。 在这个过程结束时，CNN的输入是一个矩阵，可以被视为单通道图像。他们的实验结果在DCASE 2016评估数据集的工作精度为86.2％。

Bae等人（Bae，2016）研究了长期短时记忆（LSTM）和DNN的并行组合，提出了使用顺序信息的神经网络架构。该结构由两个独立的低层网络和一个高层

网络组成。这些层分别称为LSTM层，CNN层和连接层。 LSTM层从连续的音频特征中提取连续的信息。 CNN层从谱图中学习谱时间局部性。最后，连接层汇总两个网络的输出，以便通过组合它们来利用LSTM和CNN的互补特性。RNN的核心思想是隐藏层之间的循环连接允许先前输入的内存保留内部状态，这会影响输出。 然而，在训练阶段RNN主要有两个问题需要解决：消失梯度和爆炸梯度问题（Pascanu等，2013）。 当计算反向传播过程中激励函数的导数时，长期分量可能快速指数地变为零。 这使得模型很难学习时间上遥远的输入之间的相关性。 同时，当训练期间梯度呈指数增长时，会出现爆炸梯度问题。 为了解决这个问题，提出了LSTM架构（Hochreiter等，1997）。 LSTM层由循环连接的存储块组成，其中一个存储单元包含三个乘法门。 门执行写，读和重置操作的连续类比，这使得网络能够在一段时间内利用时间信息。尽管组合的神经网络平均获得了更高的性能，但并没有给出所有场景的最佳分类结果。在巴士案中，CNN的表现优于其他网络。在公园的案例中，LSTM有更好的结果。在住宅区的情况下，DNN取得了较高的成绩。这可以被解释为所提出的网络不能完全训练一些声场，并且这些场景可能不包含足够的时间信息。未来的研究将处理更强大的网络架构，以提取声场的独特特征。提出的方法被发现可以提高分类性能，并达到79.15％的平均准确率。 DCASE 2016挑战中基于MFCC和GMM的音频场景分类任务的基线准确率为72.6％。他们的方法将性能提高了6.6％。最后，评估数据集的准确性为84.1％。

金在勋等人（kim等，2016），将深度机器的规模扩大到包括数百个网络，并将其应用于ASC。 为此，采用了几种最近的学习技术来加速训练过程，并且提出了一种新颖的随机特征多样化方法，以允许来自每个组成网络的不同贡献。为了赋予多样性，他们应用了功能明智和框架明智的方法。 作为一种特征明智的策略，他们将零相分量分析（ZCA）白化应用于预处理，并对变换权重矩阵执行随机丢弃。 程序描述如下：计算给定记录的音频信号的时频表示；样本的完成数量和参数协方差矩阵的参数；使用特征分解算法计算特征值，如奇异值分解（SVD）；选择任意一个原始基础的随机模型，然后将输入数据与选定的特征基础进行对照。所提出的方法比基线系统显示出约9％的改善，其以相对规范的方式利用MFCC和GMM。即使使用单一组成模型，它也显示出比基线更好的性能。

所提出的模型不仅能精确地分类输入的声学环境，而且表明了相对较深和较大的神经网络模型的大规模集合对稳定甚至提升模型准确性是有效的。在评估集上，所提出的集成模型达到总体准确率的85.4%，并且这个结果表明，交叉验证设置和集合方法并没有导致模型过度配合。

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。

2.3. 本文的研究方向

总结以上卷积神经网络应用在音频场景分类的实验结果，可以看出，直接将音频文件导入至卷积神经网络中，学习的效果并没有预期的好。因此需要一定的预处理（特征提取）。在特征提取部分，总结之前的研究成果，本文采用效果较好的log-mel频谱。卷积神经网络部分将引入目前广受欢迎的Tensorflow深度学习任务框架。让计算机进行音频自动分类，并探讨深度学习参数如神经元的数量，隐藏层的数量和引入丢失对学习结果造成的影响。

与之对比的基线系统，本文采用 MFCC+KNN，以对比观察卷积神经网络在音频分类中的优势。KNN 分类器是机器学习领域使用广泛的分类器，其原理简单、效果好，特别是随着训练样本数据的增多，分类效果会更好。

3. 卷积神经网络的基础理论

卷积神经网络（Convolutional Neural Network, CNN）是一种前馈神经网络，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。下面，就本文涉及到的有关卷积神经网络的理论进行论述。

3.1. 层结构

本质上，卷积神经网络是多层感知器模型的简单扩展。但是，它们的架构差异使其具有十分不同的输出。典型的卷积神经网络由多个不同的层堆叠在一起构成：一个输入层，一组可以以各种方式组合的卷积层，有限数量的完全连接的隐藏层以及一个输出层。与多层感知相比，实际差异在于引入了卷积和合并操作的组合，如图1所示。

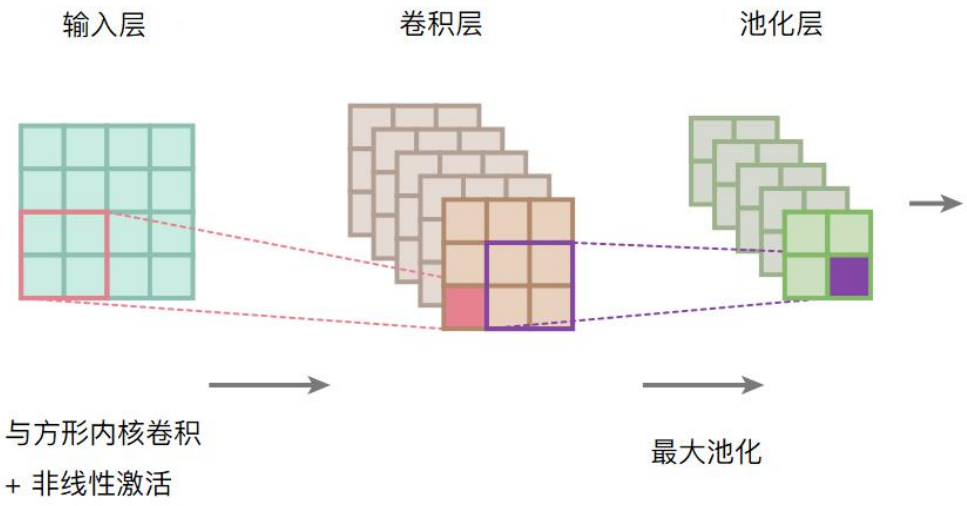

图1. 对输入数据执行的典型卷积池操作的示意图

卷积层引入了一种特殊的隐藏单元组织方式，旨在利用二维输入数据中存在的局部结构。每个隐藏单元不是连接到来自上一层的所有输入，而是仅限于处理整个输入空间的一小部分（如小的3×3像素块），称为接受域。这种隐藏单元的权重创建了一个应用于整个输入空间的卷积核，从而生成一个特征映射。这样，

一组权重可以重新用于整个输入空间。这是基于这样的前提：局部有用的特征在输入空间的其他地方也是有用的 - 这种机制不仅大大减少了要估计的参数的数量，而且提高了对数据的平移移位的稳健性。典型的卷积层将包含众多过滤器。

进一步的降维可以通过合并层来实现，合并相邻单元格的特征映射。执行的最常见的池化操作是取最大值或输入单元的平均值。这种下采样进一步提高了翻译的不变性。

3.2. 整流线性单位（ReLUs）

线性整流层（Rectified Linear Units layer, ReLU layer）使用线性整流（Rectified Linear Units, ReLU）f(x)=max(0,x)作为这一层神经的激励函数（Activation function）。它可以增强判定函数和整个神经网络的非线性特性，而本身并不会改变卷积层。

传统上，sigmoid和双曲正切函数已被用作多层感知器装置中的典型非线性激活函数。在最近的深度架构方案中，已经出现了更好的解决方案来替代它们。其中最常见的是应用整流线性单位（ReLU），它使用以下激活函数：

$$f(x) = \max(0,x) \#(3-1)$$

与传统激活函数相比，ReLUs有几个优点：更快速的计算速度和更有效的梯度传播（它们不像S形单元那样饱和），单侧生物可信度和稀疏激活结构（Glorot等，2011），尽管其构造简单但仍然保留了充分的鉴别特性。ReLUs的缺点之一是：根据随机重量初始化的状态，多个单元可能过早地落入“死区” - 输出恒定的零梯度。出于这个原因，已经提出了具有非零斜率的替代方案，例如泄漏整流线性单位（Maas等，2013），并且有实证证实了它们的有用性（Xu等，2015）。

3.3. 丢弃学习（Dropout Learning）

深度神经结构具有过度拟合的自然倾向。即使在卷积神经网络中，通过权重共享来减少参数的数量，估计值的数量大多数是训练案例数量的一个数量级。这可能会导致不良的样本外推广。

解决这个问题的一种方法是以引入丢弃学习（Hinton等，2012）。丢弃学习，即在每次训练迭代中，每个隐藏单元以预定义概率（最初为50％）被随机删除，并且学习过程正常继续。这些随机扰动有效地阻止网络学习虚假依赖，并在隐藏单元之间创建复杂的协调。这样，大群神经元不仅在其他神经元的情况下变得有用。由丢失引入的体系结构平均尝试确保每个隐藏单元学习通常有利于产生正确分类答案的特征表示。

4. 主要研究的内容

常见的基于卷积神经网络的音频分类系统的基本流程见如图 2。

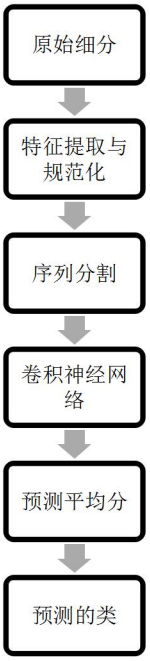

图2. 基于卷积神经网络的音频分类系统流程

分析之前研究者对基于卷积神经网络的音频分类方法的实现及该流程，现在将论述本文对音频分类的具体实现。

4.1. 特征提取部分

特征提取研究的问题是，如何从不同长度的音频中获得大小相同的片段，以及可以将哪些音频特性作为单独的通道馈入网络。 一旦有了 CNN 的初始数据集。就可以根据需要训练成深层网络。下面将介绍本文用到的部分重要特征。

Mel 频率倒谱系数（MFCC）是 ASC 中使用的最受欢迎的特征之一。它们最初是为了语音识别而开发的，它解释了 Mel-Scale 的用法，即音调的感知尺度。Mel 标度描述了人耳频率的非线性特性，它与频率的关系可用下式近似表示：

$$Mel(f) = 2959 \times \lg\left(1 + \frac{f}{700}\right) \#(3-2)$$

MFCC 是离散余弦变换，对数变换和傅立叶变换的组合，它提供了将声音激励（音高）与声道（原音）分开的可能性。这种分离使得可以识别相同的声音，特别是在源独立声音分类方面，MFCC 是一个有价值的工具。MFCC 的缺点在于，其特征为短时特征，不适合用作长时统计值。而且其对背景噪声的区分度比较不理想。

低层次的基于时域和频域的音频特征包括：过零率，它衡量信号内信号的平均变化率，并与单声道声音的主要频率有关;光谱质心，它测量光谱质量的中心，它与亮度的感知有关;以及频谱滚降，其识别频谱的幅度低于设定阈值的频率。

频率能量特征（能量/频率）：各种 ASC 系统使用的这类特征是通过对特定频段上的幅度谱或功率谱进行积分来计算的。所得到的系数测量不同子带内存在的能量的量，并且还可以表示为子带能量与编码信号中最突出的频率区域的总能量之间的比率。

4.2. 卷积神经网络部分

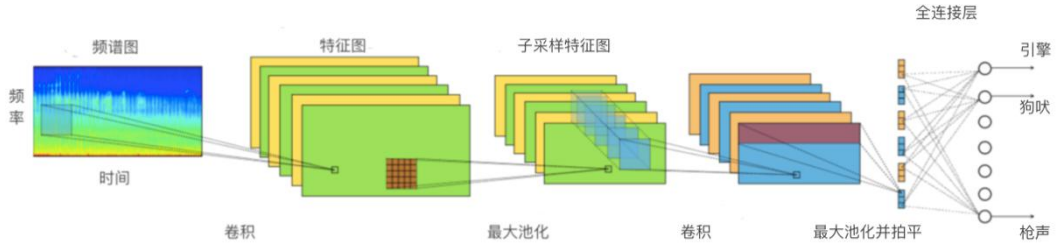

图3. 卷积神经网络部分的具体流程

在提取到音频特征之后，将得到音频的特征频谱图。第一层对输入谱图执行卷积，并在两个维度上步进。然后使用最大池化层对所获得的特征映射进行子采样，以减少图像空间的大小。第二个卷积层与第一个卷积层相同，而且使用更多的内核以获得更高级别的特征。两个卷积层使用的内核皆为整流器线性单元

（ReLU）。然后执行第二次和最后一次子采样，以针对时间轴的“破坏”。最后，由于分类涉及多个不同的类，最后一层是由全连接神经元构成的 softmax 层，以生成需要的类数量的输出。

4.3. 模型训练部分

训练的方法由两个阶段组成。第一阶段称为非全面训练，首先将整个训练数据分成两个子集：一个用于训练，另一个用于验证。在训练的整个过程中都会将训练谱图收集到逐类特征列表中，以便在序列分割之前对它们进行随机洗牌和时移。这样做是为了增加输入的可变性，因此显示网络总是略有不同的序列谱图。然后，每固定时期检查训练集和验证集上的分段性能。检查之后，如果分段验证分数得到改善，将保存网络参数。最后，如果在固定一段时间的训练后没有改善记录，将会停止训练。通过这种设置，可以发现分段验证性能容易饱和。这意味着乐谱开始以固定的稳定值振荡。当发生这种情况时，表明系统已经收敛，因此可以进入第二阶段，称之为全面培训。在这个阶段，将对固定时间的所有训练数据重新训练网络。通过查看非全面训练期间分段验证准确度的收敛时间来选择此数字。以这种方式训练的模型将达到收敛状态，没有过度的过度训练并充分利用所有可用的训练数据。

5. 已经完成的工作

进行音频分类需要样本足够的数据集，本文选取Ubransound8K音频数据集。该数据集包含来10类的城市声音：空气净化器、汽车轰鸣声、小孩玩耍声、狗吠、钻井声、发动机怠速声、枪声、手提钻声、警笛和街道杂音。共有8732个标记的声音片段，每个声音片段时长小于等于4秒。所有音频摘录摘自上传至www.freesound.org的现场录音。这些文件被预先分类为十倍个文件夹，文件夹名为fold1-fold10，以帮助比较上面的自动分类结果。除了声音片段之外，还提供了一个包含有关每个片段的元数据的CSV文件。

要进行特征提取，首先要进行特征图的绘制。这里使用基于Python的Matplotlib库（Hunter等，2007）和Librosa库（McFee等，2015）。Matplotlib的specgram

方法执行所有需要的光谱计算和绘图。Librosa提供了方便的波和对数谱图绘图方法。

现在，从Ubransound8K的fold1文件夹中抽取一些类的音频文件，以观察他们的波形差异，实现代码如下：

```
import glob
import os
import librosa
import librosa.display as ldp
import numpy as np
import matplotlib.pyplot as plt
import tensorflow as tf
from matplotlib.pyplot import specgram


def load_sound_files(file_paths):
    raw_sounds = []
    for fp in file_paths:
        X, sr = librosa.load(fp)
        raw_sounds.append(X)
    return raw_sounds

def plot_waves(sound_names,raw_sounds):
    i = 1
    fig = plt.figure(figsize=(20,20), dpi = 50)
    for n,f in zip(sound_names,raw_sounds):
        plt.subplot(8,1,i)
        ldp.waveplot(np.array(f),sr=22050)
        plt.title(n.title())
        i += 1
    plt.suptitle("Figure 1: Waveplot",x=0.5,
y=0.915,fontsize=16)
    plt.show()

def plot_specgram(sound_names,raw_sounds):
    i = 1
    fig = plt.figure(figsize=(20,20), dpi = 50)
    for n,f in zip(sound_names,raw_sounds):
        plt.subplot(8,1,i)
        specgram(np.array(f), Fs=22050)
        plt.title(n.title())
```

```
        i += 1
    plt.suptitle("Figure 2: Spectrogram",x=0.5,
y=0.915,fontsize=16)
    plt.show()

def plot_log_power_specgram(sound_names,raw_sounds):
    i = 1
    fig = plt.figure(figsize=(20,20), dpi = 50)
    for n,f in zip(sound_names,raw_sounds):
        plt.subplot(8,1,i)
        D = librosa.power_to_db(np.abs(librosa.stft(f))**2,
np.max)
        ldp.specshow(D,x_axis='time' ,y_axis='log')
        plt.title(n.title())
        i += 1
    plt.suptitle("Figure 3: Log power spectrogram",x=0.5,
y=0.915,fontsize=18)
    plt.show()

sound_file_paths =
["57320-0-0-7.wav","24074-1-0-3.wav","15564-2-0-1.wav","31323-
3-0-1.wav","46669-4-0-35.wav","89948-5-0-0.wav","40722-8-0-4.w
av"]

sound_names = ["air conditioner","car horn","children playing",
"dog bark","drilling","engine idling", "gun shot"]

raw_sounds = load_sound_files(sound_file_paths)

plot_waves(sound_names,raw_sounds)
plot_specgram(sound_names,raw_sounds)
plot_log_power_specgram(sound_names,raw_sounds)
```

生成的波形图如下：

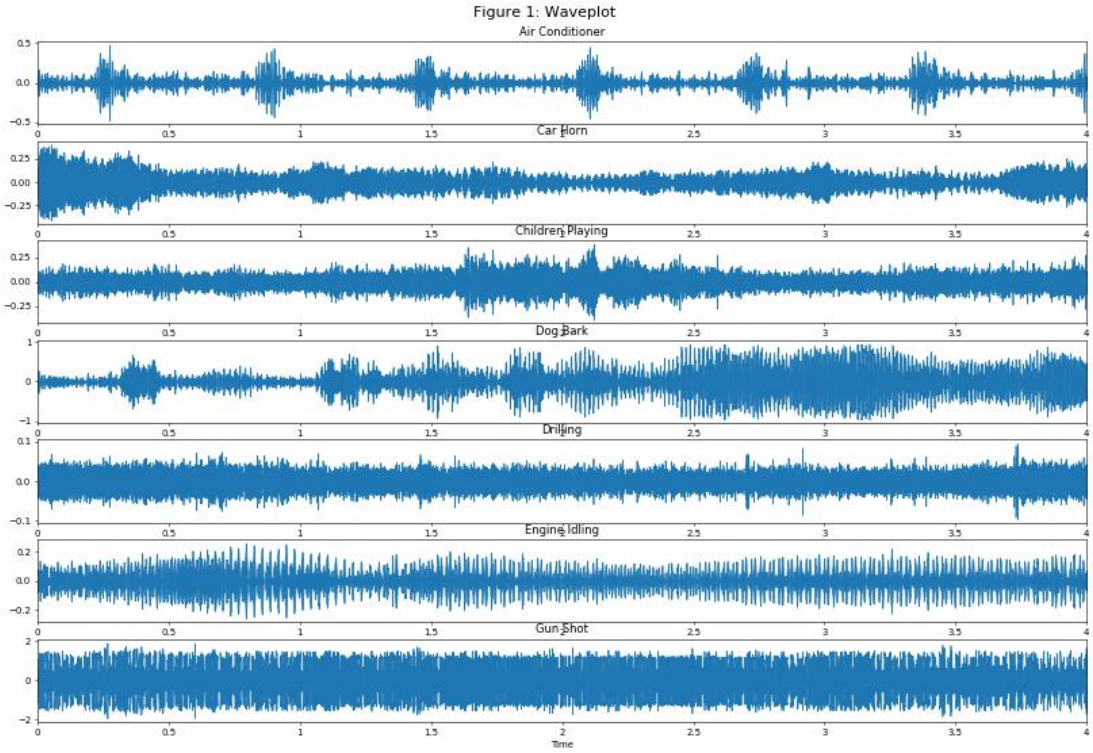
Figure 1: Waveplot
Air Conditioner
Car Horn
Children Playing
Dog Bark
Drilling
Engine Idling
Gun Shot
Time

图4. 波形图

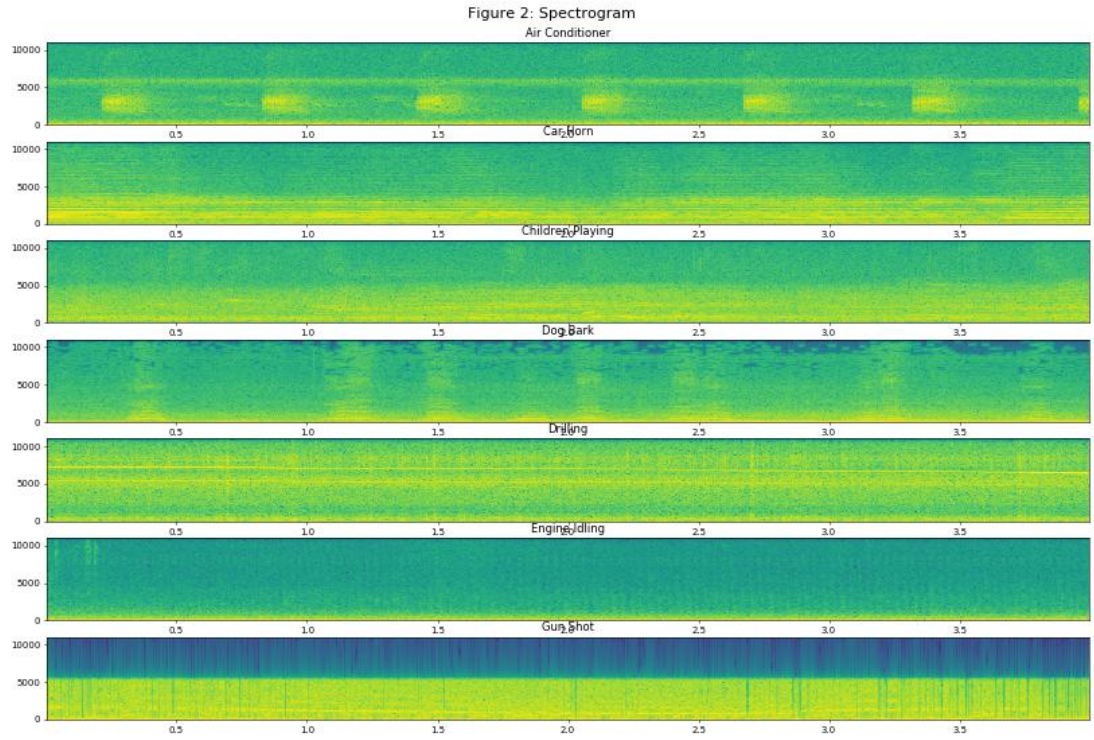
Figure 2: Spectrogram
Air Conditioner
Car Horn
Children Playing
Dog Bark
Drilling
Engine Idling
Gun Shot

图5. 频谱图

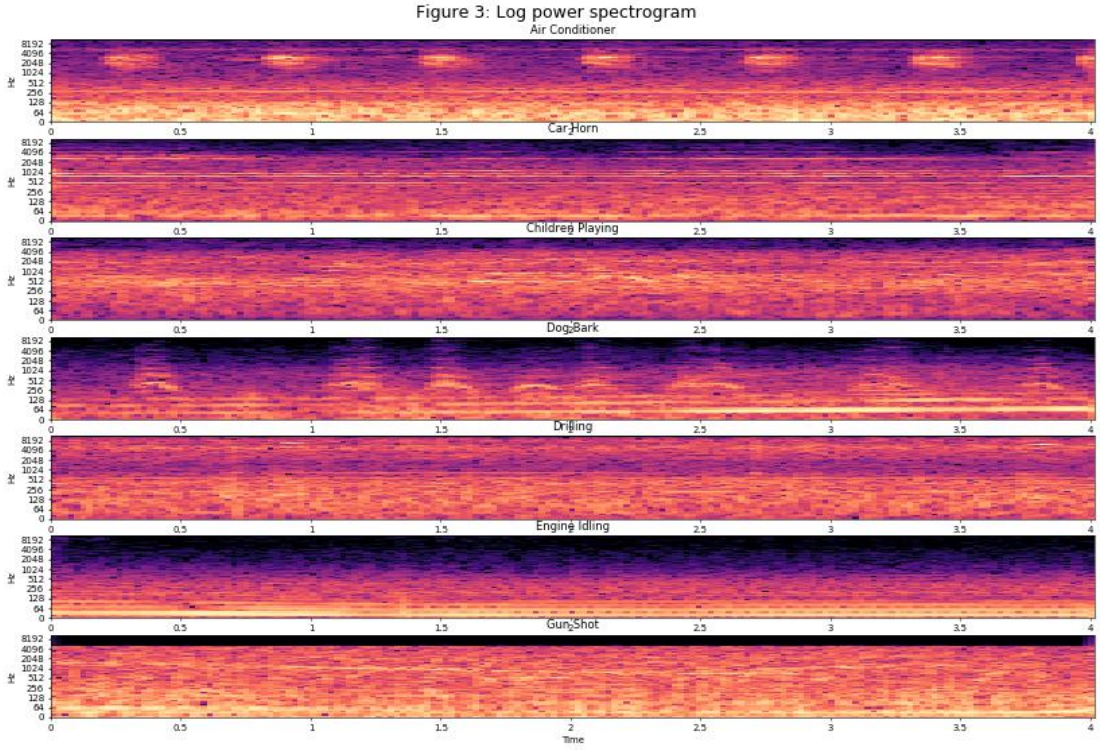

图6. 对数能量谱图

其中，波形图体现了音频信号的包络，频谱图体现了音频信号的时间-频率分布。而对数能量谱图则体现了重要的能量信息。有了这些谱图，就方面了我们从不同角度发掘音频信号的特征。

为了从声音数据中提取有用的特征，将使用 Librosa 库。它提供了几种方法从声音片段中提取不同的功能。现在将使用下面提到的方法来提取各种功能：

- melspectrogram：计算 Mel 比例功率谱图
- mfcc：梅尔频率倒谱系数
- chorma-stft：根据波形或功率谱图计算色度图
- spectral_contrast：计算光谱对比度
- tonnetz：计算音调质心特征

为了使声音剪辑的特征提取过程变得容易，定义了两个辅助方法。首先 parse_audio_files 将父目录名称，父目录中的子目录和文件扩展名（默认为.wav）作为输入。然后迭代子目录中的所有文件并调用第二个辅助函数 extract_feature。它将文件路径作为输入，通过调用 librosa.load 方法读取文件，提取并返回上面讨论的功能。这两种方法都是将原始声音片段转换为信息要素（以及每个声音片段的类别标签）所需的全部内容，我们可以将其直接输入到分类器中。每个声音片段的类别标签都在文件名中,例如，如果文件名是 108041-9-0-4.wav 那么类标签将为 9。做字符串拆分 - 并采取数组的第二项将给出类标签。特征提取的实现代码如下：

```
def extract_feature(file_name):
    X, sample_rate = librosa.load(file_name)
    stft = np.abs(librosa.stft(X))
    mfccs = np.mean(librosa.feature.mfcc(y=X, sr=sample_rate,
n_mfcc=40).T,axis=0)
    chroma = np.mean(librosa.feature.chroma_stft(S=stft,
sr=sample_rate).T,axis=0)
    mel = np.mean(librosa.feature.melspectrogram(X,
sr=sample_rate).T,axis=0)
    contrast = np.mean(librosa.feature.spectral_contrast(S=stft,
sr=sample_rate).T,axis=0)
    tonnetz =
np.mean(librosa.feature.tonnetz(y=librosa.effects.harmonic(X),
    sr=sample_rate).T,axis=0)
    return mfccs,chroma,mel,contrast,tonnetz

def parse_audio_files(parent_dir,sub_dirs,file_ext="*.wav"):
    features, labels = np.empty((0,193)), np.empty(0)
    for label, sub_dir in enumerate(sub_dirs):
        for fn in glob.glob(os.path.join(parent_dir, sub_dir,
file_ext)):
            try:
              mfccs, chroma, mel, contrast,tonnetz =
extract_feature(fn)
            except Exception as e:
              print "Error encountered while parsing file: ", fn
              continue
```

```
            ext_features =
np.hstack([mfccs,chroma,mel,contrast,tonnetz])
            features = np.vstack([features,ext_features])
            labels = np.append(labels,
fn.split('/')[2].split('-')[1])
    return np.array(features), np.array(labels, dtype = np.int)

def one_hot_encode(labels):
    n_labels = len(labels)
    n_unique_labels = len(np.unique(labels))
    one_hot_encode = np.zeros((n_labels,n_unique_labels))
    one_hot_encode[np.arange(n_labels), labels] = 1
    return one_hot_encode
parent_dir = 'Sound-Data'
tr_sub_dirs = ["fold1","fold2"]
ts_sub_dirs = ["fold3"]
tr_features, tr_labels =
parse_audio_files(parent_dir,tr_sub_dirs)
ts_features, ts_labels =
parse_audio_files(parent_dir,ts_sub_dirs)

tr_labels = one_hot_encode(tr_labels)
ts_labels = one_hot_encode(ts_labels)
```

以上，就是目前已经完成的工作，包含了特征提取需要的波形图的绘制以及特征提取的初步实现。

6. 下一步的研究内容

在通过 extract_feature 函数提取到特征以后，接下来就是实现卷积神经网络要使用到的辅助函数，如权重变量与偏执变量等。然后进入卷积阶段，通过定义一个 apply_convolution 函数进行卷积。该函数将输入数据，内核/文件大小，输入和输出深度中的许多通道或输出中的通道数量。然后，它获得权重和偏差变量，应用卷积，将偏差添加到结果中，最后应用非线性（RELU）。最大池化部分需要输入数据，内核和步长大小。

定义完卷积神经网络之后，就开始两个训练部分——非全面训练和全面训练部分。直到模型进入收敛阶段。

7. 论文工作计划表

本课题的工作计划如表 1 所示。

表 1 论文工作计划表

序号	毕业论文工作阶段要求	实习或出差地　点	起止日期	检查方式
1	搜集相关基础资料及文献、技术调研	学校	2017.10-2018.05	现场检查
2	开题设计准备及答辩	学校	2018.06	现场检查
3	进行基线系统测试	学校	2018.07-2018.09	现场检查
4	训练卷积神经网络进行音频分类，调试参数并改进系统。	学校	2018.10-2019.01	现场检查
5	完成论文初稿	学校	2019.03	现场检查
6	修改、定稿、送审	学校	2019.04	现场检查

8. 论文章节安排

第 1 章-----绪论。介绍本课题的研究目的与意义，音频场景分类行业发展现状，以及主要工作和结构安排。

第 2 章-----理论基础。介绍卷积神经网络实现的基本原理，并介绍依据卷积神经网络的音频场景分类的原理。

第 3 章-----基线系统的设计。介绍了用于对比的基线系统的设计。设计基于 MFCC+KNN 进行音频场景分类。首先分别介绍 MFCC 与 KNN 的原理，然后介绍 MFCC 与 KNN 应用在音频场景分类系统下的意义。最后，给出本系统的实现。

第 4 章-----基于卷积神经网络的音频场景分类方法的设计。首先介绍了本系统的数据集 Ubransound8K，然后逐步给出特征提取、卷积神经网络的设计以及训练的的方法。最后给出系统性能评价指标。

第 5 章-----结论。给出基线系统与基于卷积神经网络的音频场景分类方法的训练结果，并评价两种分类方法，指出其优缺点。

最后-----总结与展望。首先，总结本文所做的研究工作。然后针对这些不足，提出了一些有意义的想法和建议，为下一阶段的研究与改进提供了方向。

参考文献

Schilit B, Adams N, Want R. Context-aware computing applications[C]//Mobile Computing Systems and Applications, 1994. WMCSA 1994. First Workshop on. IEEE, 1994: 85-90.

Xu Y, Li W J, Lee K K. Intelligent wearable interfaces[M]. John Wiley & Sons, 2008.

Chu S, Narayanan S, Kuo C C J, et al. Where am I? Scene recognition for mobile robots using audio features[C]//Multimedia and Expo, 2006 IEEE International Conference on. IEEE, 2006: 885-888.

Landone C, Harrop J, Reiss J. Enabling Access to Sound Archives Through Integration, Enrichment and Retrieval: The EASAIER Project[C]//ISMIR. 2007: 159-160.

Piczak K J. Environmental sound classification with convolutional neural networks[C]//Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on. IEEE, 2015: 1-6.

Sawhney N, Maes P. Situational awareness from environmental sounds[J]. Project Rep. for Pattie Maes, 1997.

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

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

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

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

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

Eronen A, Tuomi J, Klapuri A, et al. Audio-based context awareness-acoustic modeling and perceptual evaluation[C]//Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on. IEEE, 2003, 5: V-529.

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

Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks[C]//Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. 2011: 315-323.

Maas A L, Hannun A Y, Ng A Y. Rectifier nonlinearities improve neural network acoustic models[C]//Proc. icml. 2013, 30(1): 3.

Xu B, Wang N, Chen T, et al. Empirical evaluation of rectified activations in convolutional network[J]. arXiv preprint arXiv:1505.00853, 2015.

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.

McFee B, Raffel C, Liang D, et al. librosa: Audio and music signal analysis in python[C]//Proceedings of the 14th python in science conference. 2015: 18-25.

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.

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

Lidy T, Schindler A. 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, 2016, 90: 1032-1048.

Valenti M, Diment A, Parascandolo G, et al. DCASE 2016 acoustic scene classification using convolutional neural networks[C]//Proc. Workshop Detection Classif. Acoust. Scenes Events. 2016: 95-99.

Bae S H, Choi I, Kim N S. Acoustic scene classification using parallel combination of LSTM and CNN[C]//Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016). 2016: 11-15.

Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks[C]//International Conference on Machine Learning. 2013: 1310-1318.

Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

Kim J, Lee K. Empirical study on ensemble method of deep neural networks for acoustic scene classification[C]//Proc. of IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE). 2016.

McFee B, Raffel C, Liang D, et al. librosa: Audio and music signal analysis in python[C]//Proceedings of the 14th python in science conference. 2015: 18-25.

Hunter J D. Matplotlib: A 2D graphics environment[J]. Computing in science & engineering, 2007, 9(3): 90-95.