1. 研究目的及意义（2页）

第一自然段：阐述音频分类的意义及价值

第二自然段：阐述卷积神经网络的优点

第三自然段：引出CNN对音频分类的价值

1. 音频分类方法的研究现状（6页）
   1. 音频分类方法的研究现状（提取过程）
   2. 基于CNN分类方法的研究现状（尽可能详细，最好评价）
   3. 引出自己的研究方向
2. 主要研究内容（6页）

3.1． 卷积神经网络模型简述

3.2. 基于卷积神经网络的音频分类方法原理

1. 已经完成的工作（2页）

怎么提的特征，什么样的参数有影响，比较。

1. 下一步研究内容（半页）
2. 参考文献

（15页+）

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](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。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。这一模型也可以使用[反向传播算法](https://zh.wikipedia.org/wiki/%E5%8F%8D%E5%90%91%E4%BC%A0%E6%92%AD%E7%AE%97%E6%B3%95)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。

1. 音频分类方法的研究现状
   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。

* 1. 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，但迄今为止，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MFCC特征作为网络中网络CNN架构的输入来分类音频场景。

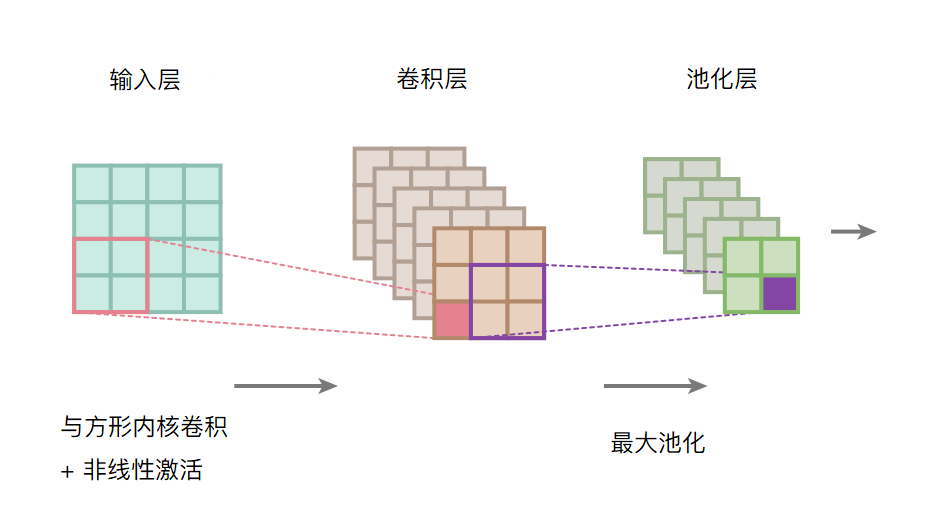
Eghbal-Zadeh等人（Eghbal-Zadeh，2016）提出了4种用于ASC的技术，即基于频谱图特征的深CNN，双耳I-矢量和CNN和I-矢量的后期融合，以提高ASC的整体准确性。在这个看不见的测试集中，我们的DCNN，BMBI，CBMBI和LFCBI准确率分别达到了83.3％，86.4％，88.7％和89.7％结果显示DCNN和BMBI方法平均执行类似。这两种方法各有优缺点。查看CBMBI结果，我们观察到校准提高了BMBI的性能。由于我们观察到DCNN和BMBI方法在不同类别上行为不同，我们期望这两者的组合改善了融合系统的性能。由此可见，LFCBI方法优于DCNN，BMBI和CBMBI系。它表明binau-ral i矢量特征和DCCN从谱图学习的表示包含关于音频场景的补充信息。因此，通过将两者结合起来，可实现了对LFCBI系统的改进。（待详细）

Hertel等人[7]提出了针对ASC的单标签分类的CNN体​​系结构和针对圆顶音频标签的DCASE 2016挑战的多标签分类。 Santoso等人[15]使用MFCC特征作为网络中网络CNN架构的输入来分类音频场景。辛德勒等人[16]通过使用常数Q变换（CQT）特征作为CNN的输入来增强结果，并且在评估数据集上实现了81.8％的准确度。他们在国内音频标签和ASC上工作。 Phan等人[17]使用CNN基于标签树嵌入（LTE）特征预先提出声学分类，并且与用于DCASE数据集声学分类的基线系统相比取得了有希望的结果。 Eghbal-Zadeh等人[18]提出了4种用于ASC的技术，即基于频谱图特征的深CNN，双耳I-矢量和CNN和I-矢量的后期融合，以提高ASC的整体准确性。 Lee等人[19]使用多倍宽度频率 - 三角洲数据增强并且在评估数据集上显示84.6％的准确度。瓦伦蒂等人[20]使用基于log mel谱图的CNN，DCASE 2016评估数据集的工作精度为86.2％。 Kim等人[21]进行了实证研究，以整合深度机器来提高ASC的性能。 Bae等人[22]研究了长期短时记忆（LSTM）和DNN的并行组合，并报告了增强的准确性。随着时间的推移，基于CNN的应用越来越受欢迎。在我们提出的方法中，我们将提出卷积神经网络和深度神经网络随机数据结构，以便与其他识别DCASE 2017数据集上的声场的方法相比，获得更准确的结果。

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. 主要研究内容
   1. 卷积神经网络模型简述
      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。

* + 1. 整流线性单位（ReLUs）

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） （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）。

* + 1. 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。

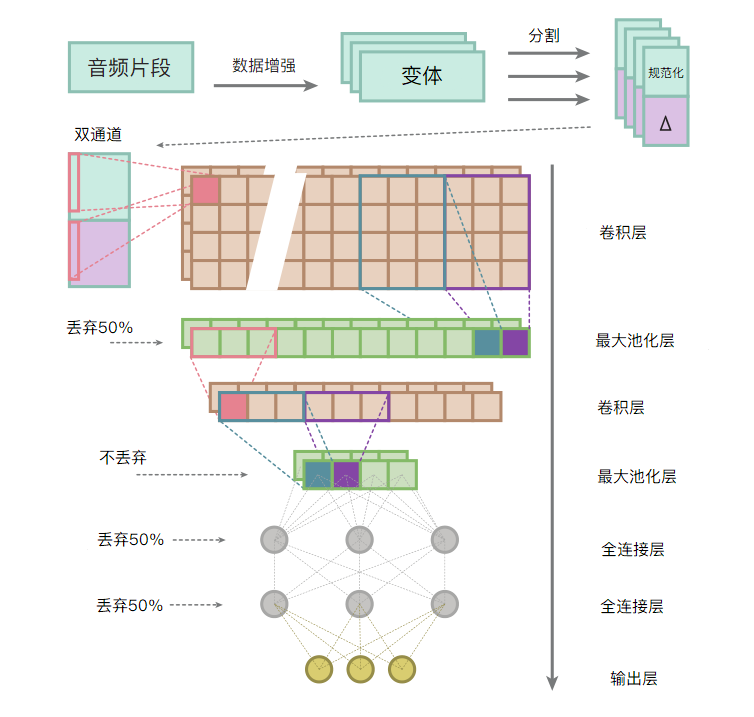
解决这个问题的一种方法是以引入丢弃学习（Hinton等，2012）。丢弃学习，即在每次训练迭代中，每个隐藏单元以预定义概率（最初为50％）被随机删除，并且学习过程正常继续。这些随机扰动有效地阻止网络学习虚假依赖，并在隐藏单元之间创建复杂的协调。这样，大群神经元不仅在其他神经元的情况下变得有用。由丢失引入的体系结构平均尝试确保每个隐藏单元学习通常有利于产生正确分类答案的特征表示。

3.2. 基于卷积神经网络的音频分类方法原理

进行音频分类需要样本足够的数据集，本文选取Ubeansound8K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](http://www.freesound.org/)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。

训练卷积神经网络涉及许多关于架构（输入数据的格式，图层的数量和大小，空间池的数量，过滤器维度）和学习超参数（学习速率，动量，批次规模，辍学概率，适用的正规化数量）。这个选择过程仍然主要基于启发式算法，特别是在进入未知领域的新应用程序时。

由于训练完整模型所需大量的时间，对所有潜在组合进行详尽评估是不可行的。因此，最现实的模型的选择必须基于对最显著因素（层数/滤波器数量，滤波器形状，学习率，丢失概率）进行的有限的单倍验证。最终系统参见图2。



**图2.** 训练系统的实现

首先，使用librosa库（McFee等，2015）对所有音频进行增强处理：重新采样到22050 Hz并归一化，提取对数坐标的梅尔频谱图。

·由于对整个剪辑的学习限制了可用于训练的示例数量，因此谱图分为41帧（短变体，约950毫秒的片段）的50％重叠片段或90％重叠的101帧（长变体，约2.3 s），丢弃过程中的无声段。

·分段（例如60行/带×41列/帧）与它们的增量（用默认librosa设置计算）一起提供为网络的双通道输入。

·第一卷积ReLU层由80个长方形（57×6尺寸，1×1步长）的滤波器组成，允许轻微的频率不变性。 Max-pooling采用4×3的泳池形状和1×3的步幅。图3给出了一小部分学习滤波器。

·第二卷积ReLU层由80个过滤器（1×3大小，1×1步幅）和最大池（1×3池大小，1×3池步幅）组成。

·通过两个完全连接的每个5000个ReLU隐藏层和一个softmax输出层进行进一步处理。

·使用pylearn2 [38]实施小批量（随机）梯度下降，即使是连续的连续批次（批量为1000），Nesterov动量为0.9 [39]，每层0.001L2重量衰减和0.5完全连接层和第一卷积层的丢失概率。

·对于短片段变体（学习率为0.002）和长时间变体（学习率为0.01）的150个时期后，停止训练过程。

·剪辑的最终预测是使用多数投票方案或者考虑每个分段预测的概率而生成的。

参考文献

1. B. Schilit, N. Adams, and R. Want. Context-aware computing applications. In Proceedings of the Workshop on Mobile Computing Systems and Applications, pages 85–90, 1994.
2. Y. Xu, W. J. Li, and K. K. Lee. Intelligent Wearable Interfaces. ISBN978-0-470-17927-7. Wiley and Sons, 2008.
3. S. Chu, S. Narayanan, C.-C. Jay Kuo, and M. J. Matari. Where am I? Scene recognition for mobile robots using audio features. In IEEE International Conference on Multimedia and Expo, pages 885–888, 2006.
4. C. Landone, J. Harrop, and J. Reiss. Enabling access to sound archives through integration, enrichment and retrieval: the easaier project. In Proceedings of the 8th International Conference on Music Information Retrieval, Vienna, Austria, September 2007.
5. 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.
6. N. Sawhney and P. Maes. Situational awareness from environmental sounds. Technical report, Massachussets Institute of Technology, 1997.
7. B. Clarkson, N. Sawhney, and A. Pentland. Auditory context awareness via wearable computing. In Proceedings Of The 1998 Workshop On Perceptual User Interfaces (PUI’98), 1998.
8. J. Ballas. Common factors in the identification of an assortment of brief everyday sounds. Journal of Experimental Psychology: Human Perception and Performance, 19(2):250–267, 1993.
9. V. T. Peltonen, A. J. Eronen, M. P. Parviainen, and A. P. Klapuri. Recognition of everyday auditory scenes: Potentials, latencies and cues. In Proceedings of the 110th Audio Engineering Society Convention, number 5404, 2001.
10. D. Dubois,C. Guastavino, and M. Raimbault.A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories. Acta Acustica united with Acustica,92:865-874,2006.
11. J. Tardieu,P. Susini,F. Poisson,P. Lazareff, and S. McAdams. Percep-tual study of soundscapes in train stations. Applied Acoustics,69:1224-1239,2008.
12. A.J. Eronen,J.T. Tuomi,A. Klapuri, and S. Fagerlund. Audio-based context awareness-acoustic modeling and perceptual evaluation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), volume 5, pages 529-532,2003.
13. Aytar Y, Vondrick C, Torralba A. Soundnet: Learning sound representations from unlabeled video[C]//Advances in Neural Information Processing Systems. 2016: 892-900.
14. X. Glorot,A. Bordes, and Y. Bengio,"Deep sparse rectifier neural networks,"in Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. JMLR W& CP,2011, vol.15, pp.315-323.
15. A.L. Maas,A.Y. Hannun, and A.Y. Ng,"Rectifier nonlinearities improve neural network acoustic models," in ICML Workshop on Deep Learning for Audio, Speech, and Language Processing,2013.
16. B. Xu,N. Wang,T. Chen, and M. Li,"Empirical evalu-ation of rectified activations in convolutional network," arXiv preprint arXiv:1505.00853,2015.
17. G.E. Hinton et al.,"Improving neural networks by preventing co-adaptation of feature detectors,"arXiv preprint arXiv:1207.0580,2012.
18. 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.
19. Eghbal-Zadeh, Hamid, et al. "CP-JKU submissions for DCASE-2016: A hybrid approach using binaural i-vectors and deep convolutional neural networks." IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) (2016).