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。

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音频场景和事件检测和分类（DCASE）挑战，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。

* 1. 基于卷积神经网络音频分类方法的研究现状

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.