卷积神经网络在音频分类中的优化

1. 音频分类
   1. 什么是音频分类[1]

想象一下，你站在火车站站台候车，你正在听周围的声音。旁边的人们喋喋不休，火车站广播正在汇报，你现在正在站台上。但是，如果有人告诉你，其实你并不在该平台，而只是在听一段来自火车站音频记录？根据你的经验，你可以判断为现在位于火车站站台。那么，电脑该如何分类这两种情况？这就是所谓的音频分类（ASC），其子类别为计算听觉场景分析领域（CASA）[22]。

* 1. 音频分类的用途

情景感知计算[2]、智能可穿戴借口[3]、移动机器人导航[4]

* 1. 音频分类的发展

早在1932年，语音识别的工作就已在贝尔实验室展开。此间花费Cherry等人20年解决鸡尾酒会问题。又过了40年，直到1990年Bregman出版了《听觉场景分析》一书后才奠基了当代该领域的研究。90年代数字助听器的发展也推动了音频分类的发展。1997年，来自MIT（麻省理工学院）的Sawhney与Maes实现了专门用于音频分类的，采用神经网络与最近邻模型的分类器。仅仅一年之后，MIT媒体实验室记录了评估样本，并提出了另一种使用隐马尔可夫模型的方法。不久之后，Eronen等人采用之前应用在语音识别上的Mel频率倒谱系数来描述音频信号的局部频谱包络，在音频分类上取得了非常好的效果。

随着音频分析领域实验数量的骤增，一个更有普适性、更大的数据库逐渐被提上历史日程。2003年，创建了视频分析的视觉检索国际权威评测（TRECVID），成为了事实标准并沿用至今。同时，IEEE音频和声学信号处理应用研讨会为其“声学场景和事件检测和分类挑战”（DCASE挑战）建立了一个新数据库，为声学场景分类中的方法设定了最先进的评估系统。

1. 深度学习

深度学习是允许由多个处理层和多个抽样层学习数据表示的计算机模型。**这些方法可以显著提高了最先进的语音识别,视觉物体识别、目标检测和许多其他领域如药物发现和基因组学等领域. 深度学习使用反向传播（BP）算法发现了在大型数据集中的复杂结构来说明如何机器应该如何改变其内部参数用于计算每一层对于上一层的表示。**深度卷积网络在图像处理、视频、语音和音频领域带来了突破，而递归网络对于连续的数据如文本和语音有很好的性能。

**深度学习在阻碍人工智能发展的问题解决方面做出了重大进步。它被证明是非常善于发现高维数据的复杂结构，因此适用于很多科学领域、商业和政府。它打破了图像识别[5-8]、语音识别[9-11]和许多机器学习领域的记录，如预测潜在药物分子的活性[12]，分析颗粒加速数据[13,14]，重建大脑回路[15]，预测非编码基因序列突变对基因表达和疾病的影响[16,17]。或许更令人惊讶的是，深度学习在自然语言理解方面的各种课题也产生了很好的效果[18]，特别是主题分类，情感分析，自动问答[19]和语言翻译[20,21]。**

**我们认为在不远的而将来深度学习将取得更大的成功，因为其需要很少的人工干预，可以很容易的利用计算量和数据的增长。目前正在为深层神经网络而开发的新的学习算法和结构只会加快这一进程。**

1. 卷积神经网络

3.1. 什么是卷积神经网络

卷积神经网络**被设计用来处理多重矩阵的数据，例如，一个彩色图像由3个2D矩阵组成。许多数据形态是多重矩阵的形式：1D信号和序列包括语音，2D图像或声谱图，3D视频或立体图像。卷积网络利用4个关键点来利用自然信号的属性：局部连接，权值共享，池化和多层结构。如图2，一个卷积网络的内部**

**卷积层的功能是检测前一层的特征的局部连接，池化层的作用是是合并特征相似的地方。这是因为形成一个主题的特征的相对位置会有所不同。一个典型的池化单元计算一个局部块的最大值，邻近的池化单元按照一行或一列或者更多的方式移动的从局部块中获取数据，从而减少了维数，增加了移动或扭曲的不变性。2到3个卷积层，非线性函数和池化层堆叠，再连接上全连接层，通过bp算法训练所有过滤器的权重，就得到了**卷积神经网络**。**

**深度神经网络开发自然信号层级组成的特性，其中，高水平的特征由低水平的组合获得。在图像中，边缘的局部组合形成图案，图案构成部分，部分组成目标。相似的结构存在于语音和文本中，如电话里的声音，音位，音节，文本里的单词和句子。当前一层的元素的位置或外貌变化时，池化能保证表示几乎不变。**

**卷积网络中卷积层和池化层是由视觉神经科学中的简单细胞和复杂细胞的经典观念启发得到的，这种细胞是以LGN–V1–V2–V4–IT这种形式形成视觉回路的。当被给予相同图片时，卷积网络可以解释猴子的下颞叶皮质的随机160个神经元的变化。卷积网络有神经认知的根源，有相似的结构，但神经认知中没有类似bp算法之类的端到端的监督学习算法。一个原始的1D卷积网络称作时延神经网络被用于音素和简单单词的识别。**

**追溯到1990s早期，卷积网络已经有了大量的应用，开始是时延神经网络用于语音识别和文本阅读。文本阅读系统使用一个训练好的卷积网络和一个而受到语言约束的概率模型的二联合。到1990s晚期，这个系统阅读美国所有支票的10%。大量基于卷积网络的视觉特征识别和手写体识别被Microsoft开发。在1990s早期，卷积网络也被实验在人脸，手写图像的目标检测和人脸识别。**

3.2. 卷积神经网络的应用

卷积神经网络可以追溯到20世纪80年代，但直到最近才被采纳为一种进行对象分类任务的方法。2012年Krizhevsky等人的工作[23]，标志着大规模视觉识别的转折点[24]。此后，代替了依靠手动标记特征的技术，卷积神经网络在众多模式识别领域都有重大进展。包括分类交通标志[25]，门牌号[26,27]和手写数字[28]，行人检测[29]和电子显微镜图像处理[30]。

尽管主要用于视觉识别环境，卷积架构也被成功应用在演讲[31-37]和音乐分析[38,39]。这些成功案例表明，发掘数据本身蕴含的信息同样可为其他领域的问题提供可行的解决方案。

1. 为什么要用深度学习解决音频分类

此前音频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。在这方面深度学习技术的引入在过去年几年中已经开始慢慢展开。

2015年，Karol J. Piczak等人[40]对深度学习中的卷积神经网络是否可有效的应用于音频分类这一问题进行了探讨。

以监督方式训练深度神经构架的主要问题之一是有效学习所需的计算量和标记数据量。虽然前者在某些部分通过硬件改进和通用GPU计算在全球范围内解决，但后者是非常依赖于数据库。为此他们选取了3个数据库用于学习：ESC-50 ，ESC-10和UrbanSound8K。

系统设计如下：现将音频数据段进行增强处理，再依据特征分组，输入卷积神经网络处理。该模型以5倍（ESC-10和ESC-50）和10倍（UrbanSound8K）交叉验证方案进行评估，单次训练折叠用作间歇验证集。

在所有情况下，基于卷积神经网络的模型都比使用手动设计的特征的各自实现表现更好，尤其是在对来自更多不同类别的ESC-50数据集的音频事件进行分类时（基线准确度：44％，最佳网络：64.5％）。

以上实验表明了使用卷积神经网络进行音频分类是一个切实可行的办法。卷积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。

1. 研究目标与内容

目标：使用当前流行的深度学习框架Tensorflow，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。研究如何从不同长度的音频片段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送入网络。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。

本系统使用以下库或框架

* Tensorflow（深度学习框架，用卷积神经网络进行分类）
* Librosa（音频信号的预处理、特征提取）
* Numpy（提供适合卷积神经网络运算的数据结构）
* Matplotlib（绘制实验数据图）
* Ubransound8K (音频片段库)

流程如下:



TensorFlow是用于表示机器学习算法的接口，以及用于执行这种算法的实现。TensorFlow的兼容性很好，可以几乎不加修改的运行在各种平台上，范围从诸如手机和平板电脑之类的移动设备，到数百台机器的大型分布式系统以及诸如GPU卡之类的数千个计算设备。该系统非常灵活，可以用来表达各种各样的算法，包括深度神经网络模型的训练和推理算法，它已被用于开展研究并将机器学习系统部署到十几个生产环境中包括语音识别，计算机视觉，机器人，信息检索，自然语言处理，地理信息提取和计算药物发现等计算机科学领域和其他领域。

参考文献

1. Gerasch M C. Acoustic Scene Classification[J]. 2015.
2. B. Schilit, N. Adams, and R. Want, “Context-aware computing applications,” in Mobile Computing Systems and Applications, 1994. WMCSA 1994. First Workshop on. IEEE, 1994, pp. 85–90.
3. Y. Xu, W. J. Li, and K. K. Lee, Intelligent wearable interfaces. John Wiley & Sons, 2008.
4. LeCun Y, Bengio Y, Hinton G. Deep learning[J]. nature, 2015, 521(7553): 436.
5. Krizhevsky, A., Sutskever, I. & Hinton, G. ImageNet classification with deep convolutional neural networks. In Proc. Advances in Neural Information Processing Systems 25 1090–1098 (2012).
6. Farabet, C., Couprie, C., Najman, L. & LeCun, Y. Learning hierarchical features for scene labeling. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1915–1929 (2013).
7. Tompson, J., Jain, A., LeCun, Y. & Bregler, C. Joint training of a convolutional network and a graphical model for human pose estimation. In Proc. Advances in Neural Information Processing Systems 27 1799–1807 (2014).
8. Szegedy, C. et al. Going deeper with convolutions. Preprint at http://arxiv.org/ abs/1409.4842 (2014).
9. Mikolov, T., Deoras, A., Povey, D., Burget, L. & Cernocky, J. Strategies for training large scale neural network language models. In Proc. Automatic Speech Recognition and Understanding 196–201 (2011).
10. Hinton, G. et al. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Processing Magazine 29, 82–97 (2012).
11. Sainath, T., Mohamed, A.-R., Kingsbury, B. & Ramabhadran, B. Deep convolutional neural networks for LVCSR. In Proc. Acoustics, Speech and Signal Processing 8614–8618 (2013).
12. Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E. & Svetnik, V. Deep neural nets as a method for quantitative structure-activity relationships. J. Chem. Inf. Model. 55, 263–274 (2015).
13. Ciodaro, T., Deva, D., de Seixas, J. & Damazio, D. Online particle detection with neural networks based on topological calorimetry information. J. Phys. Conf. Series 368, 012030 (2012).
14. Kaggle. Higgs boson machine learning challenge. Kaggle https://www.kaggle. com/c/higgs-boson (2014).
15. Helmstaedter, M. et al. Connectomic reconstruction of the inner plexiform layer in the mouse retina. Nature 500, 168–174 (2013).
16. Leung, M. K., Xiong, H. Y., Lee, L. J. & Frey, B. J. Deep learning of the tissue-regulated splicing code. Bioinformatics 30, i121–i129 (2014).
17. Xiong, H. Y. et al. The human splicing code reveals new insights into the genetic determinants of disease. Science 347, 6218 (2015).
18. Collobert, R., et al. Natural language processing (almost) from scratch. J. Mach. Learn. Res. 12, 2493–2537 (2011).
19. Bordes, A., Chopra, S. & Weston, J. Question answering with subgraph embeddings. In Proc. Empirical Methods in Natural Language Processing http:// arxiv.org/abs/1406.3676v3 (2014).
20. Jean, S., Cho, K., Memisevic, R. & Bengio, Y. On using very large target vocabulary for neural machine translation. In Proc. ACL-IJCNLP http://arxiv.org/ abs/1412.2007 (2015).
21. Sutskever, I. Vinyals, O. & Le. Q. V. Sequence to sequence learning with neural networks. In Proc. Advances in Neural Information Processing Systems 27 3104–3112 (2014).
22. Geiger J T, Schuller B, Rigoll G. Large-scale audio feature extraction and SVM for acoustic scene classification[C]//Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on. IEEE, 2013: 1-4.
23. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
24. Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.
25. CireşAn D, Meier U, Masci J, et al. Multi-column deep neural network for traffic sign classification[J]. Neural networks, 2012, 32: 333-338.
26. Goodfellow I J, Bulatov Y, Ibarz J, et al. Multi-digit number recognition from street view imagery using deep convolutional neural networks[J]. arXiv preprint arXiv:1312.6082, 2013.
27. Sermanet P, Chintala S, LeCun Y. Convolutional neural networks applied to house numbers digit classification[C]//Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012: 3288-3291.
28. Ciregan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification[C]//Computer vision and pattern recognition (CVPR), 2012 IEEE conference on. IEEE, 2012: 3642-3649.
29. Sermanet P, Kavukcuoglu K, Chintala S, et al. Pedestrian detection with unsupervised multi-stage feature learning[C]//Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013: 3626-3633.
30. Ciresan D, Giusti A, Gambardella L M, et al. Deep neural networks segment neuronal membranes in electron microscopy images[C]//Advances in neural information processing systems. 2012: 2843-2851.
31. Lee H, Pham P, Largman Y, et al. Unsupervised feature learning for audio classification using convolutional deep belief networks[C]//Advances in neural information processing systems. 2009: 1096-1104.
32. Deng L, Abdel-Hamid O, Yu D. A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion[C]//Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013: 6669-6673.
33. Abdel-Hamid O, Mohamed A, Jiang H, et al. Convolutional neural networks for speech recognition[J]. IEEE/ACM Transactions on audio, speech, and language processing, 2014, 22(10): 1533-1545.
34. Sainath T N, Mohamed A, Kingsbury B, et al. Deep convolutional neural networks for LVCSR[C]//Acoustics, speech and signal processing (ICASSP), 2013 IEEE international conference on. IEEE, 2013: 8614-8618.
35. Abdel-Hamid O, Mohamed A, Jiang H, et al. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition[C]//Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, 2012: 4277-4280.
36. Abdel-Hamid O, Deng L, Yu D. Exploring convolutional neural network structures and optimization techniques for speech recognition[C]//Interspeech. 2013, 2013: 1173-5.
37. Deng L, Li J, Huang J T, et al. Recent advances in deep learning for speech research at Microsoft[C]//Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013: 8604-8608.
38. Dieleman S, Brakel P, Schrauwen B. Audio-based music classification with a pretrained convolutional network[C]//12th International Society for Music Information Retrieval Conference (ISMIR-2011). University of Miami, 2011: 669-674.
39. Van den Oord A, Dieleman S, Schrauwen B. Deep content-based music recommendation[C]//Advances in neural information processing systems. 2013: 2643-2651.
40. 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.