PANNs：大规模预训练音频神经网络

音频模式识别

孔秋强，IEEE学生会员，曹寅，IEEE会员，Turab Iqbal，王宇轩，王文武，IEEE高级会员，Mark D.Plumbley，IEEE研究员

摘要音频模式识别是机器学习领域的一个重要研究课题，包括音频标记、声学场景分类、音乐分类、语音情感分类和声音事件检测等多项任务。最近，神经网络已被应用于解决音频模式识别问题。然而，以前的系统是建立在持续时间有限的特定数据集上的。最近，在计算机视觉和自然语言处理中，对大规模数据集进行预训练的系统已经很好地推广到了几个任务中。然而，关于大规模数据集上用于音频模式识别的预训练系统的研究有限。本文提出了在大规模AudioSet数据集上训练的预训练音频神经网络（PANNs）。这些PANN被转移到其他与音频相关的任务中。我们研究了由各种卷积神经网络建模的PANN的性能和计算复杂度。我们提出了一种称为Wavegram-Logmel-CNN的架构，该架构使用log-mel频谱图和波形作为输入特征。我们最好的PANN系统在AudioSet标记上实现了最先进的平均精度（mAP）0.439，优于之前最好的0.392。我们将PANN转移到六个音频模式识别任务中，并在其中几个任务中展示了最先进的性能。我们已经发布了PANN的源代码和预训练模型：https://github.com/qiuqiangkong/audioset_tagging_cnn.

索引术语音频标记、预训练音频神经网络、迁移学习。

一.导言

音频模式识别是机器学习领域的一个重要研究课题，在我们的生活中起着重要作用。我们被声音包围着，这些声音包含了我们所处位置的丰富信息，以及我们周围发生的事件。音频模式识别包含几个任务，如音频标记[1]、声学场景分类[2]、音乐分类[3]、语音情感分类和声音事件检测[4]。

近年来，音频模式识别引起了越来越多的研究兴趣。早期音频模式识别工作

Q.Kong、Y.Cao、T.Iqbal和M.D.Plumbley在英国Guild-ford GU2 7XH萨里大学视觉、语音和信号处理中心工作（电子邮件：q.kong@surrey.ac.uk；yin.cao@surrey.ac.uk；t.iqbal@surrey.ac.uk；m.plumbley@surrey.ac.uk）。

这项工作部分得到了EPSRC赠款EP/N014111/1《理解声音》的支持，部分得到了中国国家留学基金管理委员会研究奖学金201406150082的支持，另一部分得到了EP/N509772/1赠款下EPSRC博士培训伙伴关系的助学金（参考号：1976218）的支持。本研究得到了国家自然科学基金（11804365）的资助。（孔秋强为第一作者。）（尹曹为通讯

Y.Wang就职于美国加利福尼亚州山景城字节跳动人工智能实验室（电子邮件：wangyuxuan.11@bytedance.com）。

W.Wang来自英国吉尔福德GU2 7XH萨里大学视觉、语音和信号处理中心，以及中国青岛科技大学，邮编266071（电子邮件：w.wang@surrey.ac.uk）.

专注于个人研究人员收集的私人数据集[5][6]。例如，Woodard[5]应用隐马尔可夫模型（HMM）对三种类型的声音进行分类：木门打开和关闭、金属掉落和倒水。最近，声学场景和事件的检测和分类（DCASE）挑战系列[7][8][9][2]提供了公开可用的数据集，如声学场景分类和声音事件检测数据集。DCASE挑战引起了人们对音频模式识别越来越多的研究兴趣。例如，最近的DCASE 2019挑战赛在五个子任务中收到了311个参赛作品[10]。

然而，当在大规模数据集上训练时，音频模式识别系统的性能如何仍然是一个悬而未决的问题。在计算机视觉中，已经使用大规模ImageNet数据集构建了几个图像分类系统[11]。在自然语言处理中，已经使用维基百科等大规模文本数据集构建了几种语言模型[12]。然而，在大规模音频数据集上训练的系统更为有限[1][13][14][15]。

音频模式识别的一个里程碑是AudioSet的发布[1]，这是一个包含527个声音类别的5000多小时录音的数据集。AudioSet没有发布原始录音，而是发布了从预训练卷积神经网络中提取的音频片段的嵌入特征[13]。一些研究人员研究了具有这些嵌入特征的建筑系统[13][16][17][18][19][20]。然而，嵌入特征可能不是音频记录的最佳表示，这可能会限制这些系统的性能。在这篇文章中，我们提出了使用各种神经网络在原始AudioSet录音上训练的预训练音频神经网络（PANN）。我们发现，几个PANN系统的性能优于以前最先进的音频标记系统。我们还研究了PANN的音频标记性能和计算复杂性。

我们建议将PANN转移到其他音频模式识别任务中。之前的研究人员已经研究了音频标记的迁移学习。例如，在[21]中提出的百万首歌曲数据集上对音频标记系统进行了预训练，从预训练的卷积神经网络（CNN）中提取的嵌入特征被用作第二阶段分类器的输入，如在MagnaTagATune上预训练的神经项[23]，声学场景[24]数据集在其他音频标记任务上进行了微调[25][26]。这些迁移学习系统主要使用音乐数据集进行训练，并且仅限于比AudioSet更小的数据集。

networks or support vector machines (SVMs) [14][22]. Sys-

这项工作的贡献包括：（1）我们介绍