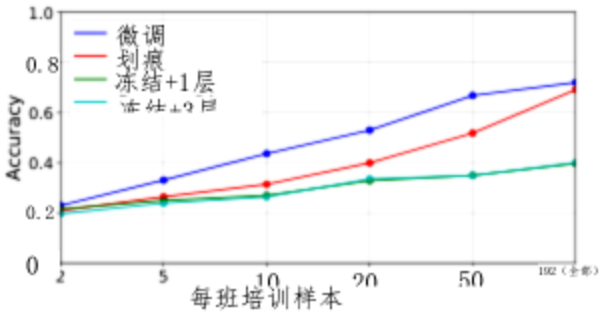

图11。每节课有不同数量的训练片段，RAVDESS的准确性。

表十八
雷达的准确性

	斯托阿[58]	划痕	微调	免费 L1	Freeze_L3
Acc。	0.645	0.692	0.721	0.397	0.401

24名专业演员，包括12名女性和12名男性，模拟8种情绪，如快乐和悲伤。任务是将每个声音片段分类为一种情感。开发集中有1440个音频片段。我们通过4倍交叉验证来评估我们的系统。表XVIII显示，曾等人[58]提出的先前最先进的系统达到了0.645的精度。我们从头开始训练的CNN14系统达到了0.692的精度。微调的CNN14系统达到了0.721的最先进精度。Freeze_L1和Freeze_L3系统分别达到了0.397和0.401的较低精度。图11显示了系统相对于一系列训练片段的准确性。微调的系统和从头开始训练的系统。其性能优于使用PANN作为特征提取器的系统。这表明RAVDESS数据集的录音可能具有不同的AudioSet数据集分布。因此，需要对PANN的参数进行微调，以在RAVDESS分类任务中实现良好的性能。

E. 讨论

在这篇文章中，我们研究了用于AudioSet标记的各种PANN。我们提出的几个PANN的表现优于之前最先进的AudioSet标记系统，包括CNN14的mAP达到0.431，ResNet38的mAP为0.434，优于谷歌的0.314基线。MobileNets是轻量级的系统，具有较少的多重加法和参数数量。MobileNetV1的mAP达到了0.389。我们改进的一维系统Res1dNet31的mAP为0.365，优于之前的一维CNN，包括0.295的DaiNet[31]和0.266的LeeNet11[42]。我们提出的Wavegram-Logmel CNN系统在所有PANN中实现了最高的0.439 mAP。PANNs可以用作新音频模式识别任务的预训练模型。

在AudioSet数据集上训练的PANN被转移到六个音频模式识别任务中。我们证明，经过微调的PANN在ESC-50、MSOS和RAVDESS分类任务中实现了最先进的性能，并接近在DCASE 2018任务2和GTZAN分类任务中取得了最先进的性能。在PANN系统中，经过微调的PANN在新任务上的表现总是优于从头开始训练的PANN。实验表明，PANNs在有限训练数据的情况下成功地推广到其他音频模式识别任务。。

七、结论

我们提出了在AudioSet上训练的预训练音频神经网络（PANNs），用于音频模式识别。人们研究了各种神经网络来构建PANN。我们提出了一种从波形中学习的波形图特征，以及一种在AudioSet标记中实现最先进性能的波形图Logmel CNN，存档了0.439的mAP。我们还研究了PANN的计算复杂性。我们证明，PANN可以转移到广泛的音频模式识别任务中，并优于之前几种最先进的系统。当对新任务的少量数据进行微调时，PANN可能很有用。未来，我们将把PANN扩展到更多的音频模式识别任务。

参考文献

[1] J.F.Gemmeke，D.P.Ellis，D.Freedman，A.Jansen，W.Lawrence，R.C。Moore，M.Plakal和M.Ritter，*音频集：本体论和人类-标记音频事件的标记数据集”，IEEE国际会议声学、语音和信号处理（ICASSP），2017，第776780页。

[2] A.Mesaros、T.Heittola和T.Virtanen，一个多设备数据集城市声学场景分类”，在探测和声学场景和事件分类（DCASE），2018，第9页13.

[3] K.Choi、G.Fazekas和M.Sandler，“使用深度自动标记卷积神经网络，”在国际会议音乐信息检索学会（ISMIR），2016，第805811页。

[4] E.Cakir、T.Heittola、H.Huttunen和T.Virtanen，复调声音国际上使用多标签深度神经网络的事件检测神经网络联合会议（IJCNN），2015年。

[5] J.P.Woodard，按产品分类的自然声音建模和分类代码隐马尔可夫模型、IEEE信号处理学报，1992年，第40卷，第18331835页。

[6] D.P.W.Ellis，检测警报声，https://academiccommons.哥伦比亚.edu/doi/10.7916/D8F19821/下载，2001年。

[7] D.斯托维尔、D.吉安努利斯、E.贝内托斯、M.拉格朗日和M.D。Plumbley，声学场景和事件的检测和分类，IEEE多媒体汇刊，第17卷，第173317462015页。

[8] A.梅萨罗斯、T.海托拉、E.贝内托斯、P.福斯特、M.拉格朗日、T.维尔塔宁。和M.D.Plumhlev，*声学场景的检测和分类和活动：2016年DCASE挑战赛的结果，IEEE/ACM音频、语音和语言处理学报（TASLP），第26卷，第379393页，2018年。

[9] A.梅萨罗斯、T.海托拉、A.迪蒙特、B.伊丽莎白、A.沙阿、E.文森特，B.Raj和T.Virtanen，DCASE 2017挑战设置：任务、数据集以及基线系统”，在关于检测和分类的研讨会上声学场景与事件（DCASE），2017，第8592页。

[10] 2019年DCASE挑战赛，http://dcase.community/challenge2019，2019.

[11] 邓、董、索彻、李、李、费飞，ImageNet：IEEE大会上的大规模分层图像数据库计算机视觉与模式识别（CVPR），2009，第248255页。

[12] J.Devlin，M.-W.Chang，K.Lee和K.Toutanova，BERT：预-训练用于语言理解的深层双向转换器，协会北美分会年会计算语言学（NAACL），2018，第41714186页。

[13] S.好时、S.乔杜里、D.P.埃利斯、J.F.杰梅克、A.詹森、R.C。Moore、M.Plakal、D.Platt、R.A.Saurous、B.Seybold等人，美国有线电视新闻网IEEE国际标准中的“大规模音频分类架构”2017年声学、语音和信号处理会议（ICASSP），第131135页。