在AudioSet上训练的PANN有190万个音频片段，包含527个声音类别的本体；（2） 我们研究了各种PANN的音频标记性能和计算复杂度之间的权衡；（3） 我们提出了一个我们称之为Wavegram-Logmel CNN的系统，该系统在AudioSet标记上的平均精度（mAP）为0.439，优于之前最先进的mAP 0.392系统和谷歌的mAP 0.314系统；（4） 我们证明，PANN可以转移到其他音频模式识别任务中，优于几种最先进的系统；（5） 我们已经发布了源代码和预训练的PANN模型。

本文的结构如下：第二节介绍了使用各种卷积神经网络的音频标记；第三节介绍了我们提出的Wavegram CNN系统；第四节介绍了我们的数据处理技术，包括AudioSet标签的数据平衡和数据增强；第六节显示了实验结果，第七节总结了这项工作。。

二、音频标签系统

音频标记是音频模式识别的一项重要任务，其目标是预测是否存在。音频剪辑中的音频标签。音频标记的早期工作包括使用手动设计的特征作为输入，如音频能量、过零率和梅尔频率倒谱系数（MFCC）[27]，生成模型，包括高斯混合模型（GMMs）[28][29]、隐马尔可夫模型（HMM）和判别支持向量机（SVM）[30]已被用作分类器。最近，基于神经网络的方法，如卷积神经网络（CNN），已被用于预测录音的标签[3]。基于CNN的系统在几个DCASE挑战任务中取得了最先进的性能，包括声学场景分类[2]和声音事件检测[4]。然而，这些作品中的许多都集中在声音类别数量有限的特定任务上，并且不是为了识别各种各样的声音类别而设计的。在这篇文章中，我们专注于在AudioSet[1]上训练大规模PANNs，以解决一般的音频标记问题。

A. CNN

1） 传统的卷积神经网络：卷积神经网络已成功应用于计算机视觉任务，如图像分类[31][32]。CNN由几个卷积层组成，每个卷积层包含几个与输入特征图卷积的核，以捕获它们的局部模式。用于音频标记的CNN[3][20]通常使用log-mel频谱图作为输入[3][20]。短时傅里叶变换（STFT）应用于时域波形以计算频谱图。然后，将梅尔滤波器组应用于频谱图，然后进行对数运算以提取对数梅尔频谱图[3][20]。

2） 使CNN适应AudioSet标签：我们使用的PANN基于我们之前为DCASE 2019挑战提出的时间任务CNN系统[33]，并在CNN的倒数第二层添加了一个额外的全连接层

表一

用于音频设备标记的CNNS

VGGish [1]	CNN6	CNN10	CNN14
对数梅尔光谱仪 ram 96 frames × 64 mel bins	对数梅尔光谱仪 n 1000 frames × 64 mel bins		
3 3 64 ReLU	5 5 @ 64 BN, ReLU	(3 × 3 @ 64 BN, ReLU) × 2	(3 364 BN, ReLU) × 2
MP 2 × 2	Pooling 2 × 2		
3 3 128 ReLU	5 × 5 @ 128 BN, ReLU	(3 3 128 BN, ReLU) × 2	(3 3 128 BN, ReLU) × 2
MP 2 2	Pooling 2 × 2		
(3 × 3 @ 256 ReLU) 2.	5 × 5 @ 256 BN, ReLU	(3 × 3 @ 256 BN, ReLU) × 2	(3 3克256 BN, ReLU) × 2
MP 2 × 2	Pooling 2 × 2		
(3 × 3 @ 512 ReLU) × 2	5 5 512 BN, ReLU	(3 × 3 @ 512 BN, ReLU) × 2	(3 3 512 BN, ReLU) 2
MP 2 × 2 平坦 ，	全球汇集		泳池2 2
FC 4096 ReLU × 2	FC 512, ReLU		(3 3 @ 1024 BN, ReLU) × 2
FC 527, Sigmoid	FC 527, Sigmoid		Pooling 2 × 2
			(3 × 3 @ 2048 BN, ReLU) × 2
			全球的 pooling
			FC 2048, ReLU
			F527签名 noid

以进一步提高表现能力。我们研究了6-10层和14层CNN。基于AlexNet[34]，6层CNN由4个卷积层组成，核大小为5.5。10层和14层CNN分别由4层和6层卷积层组成，灵感来自类似VGG的CNN[35]。每个卷积块由2个卷积层组成，核大小为3.3。在每个卷积层之间应用批归一化[36]，并使用ReLU非线性[37]来加速和稳定训练。我们对每个卷积块应用大小为2 2的平均池进行下采样，因为2 2平均池已被证明优于2 2最大池[33]。

在最后一个卷积层之后应用全局池，将特征图汇总为固定长度的向量。在[15]中，全局池化使用了最大和平均操作。为了结合它们的优点，我们将平均向量和最大向量相加。在我们之前的工作[33]中，这些固定长度的向量被用作音频片段的嵌入特征。在这项工作中，我们在固定长度向量上添加了一个额外的全连接层来提取嵌入特征，这可以进一步提高它们的表示能力。对于特定的音频模式识别任务，线性分类器应用于嵌入特征，然后是用于分类任务的softmax非线性或用于标记任务的sigmoid非线性。在每次下采样操作和完全连接的层之后应用Dropout[38]，以防止系统过度拟合。表一总结了我们提出的CNN系统。“@符号后的数字表示特征图的数量。第一列显示了[13]提出的VGGish网络。MP是最大池化的缩写。表I中的“池化2 2”是大小为2 2的平均池化。在[13]中，一个音频片段被分割成1秒的片段，[13]还假设每个片段都继承了音频片段的标签，这可能会导致标签不正确。相比之下，我们的系统从表I中的第二列到第四列应用了整个音频片段进行训练，而没有将音频片段分割成片段。