表二

用于音频集标记的资源网

ResNet22	ResNet38	ResNet54
对数梅尔谱图1000帧64梅尔箱		
$(3 \times 3 @ 512, \mathrm{BN}, \mathrm{ReLU}) \times 2$		
池2 2		
基础B 64 2	基础B 64 3	瓶颈B 64 3
池2 2		
基础B 128 2	基础B1284	瓶颈B 128 4
池2 2		
基础B 256 2	基础B 256 6瓶装B 256 6	
池2 2		
基础B 512 2	基础B 512 3瓶装B 512 3	
池2	× 2	
（32048，BN，ReLU）2		
全球汇集		
FC 2048，ReLU		
FC 527，Sigmoid		

表三

用于音频集标记的移动网络

MobileNetV1	MobileNetV2
$3 \times 3 @ 32$, BN, ReLU	
池2 2	
V1区块@64 V1Block@128	V2块，t=1@16 （V2块，t=6@24）2
池2 2	
V1Block@128 V1区块@256	（V2块，t=6@32）3
池2 2	
V1区块@256 V1区块@512	（V2块，t=6@64）4
池2 2	
（V1区块@512）5 V1区块@1024	（V2块，t=6@96）3
池2 2	
V1区块@1024	V2区块，t=6160）3
	(V2Block, t=6 @ 320) × 1
全球汇集	
FC，1024，ReLU	
FC，527，Sigmoid	

我们将音频片段的波形表示为xn，其中n是音频片段的索引，f（n）[0,1]K是表示K个声音类别存在概率的PANN的输出。xn的标记表示为yn{0,1}K。使用二进制交叉熵损失函数l来训练PANNs：

$$l = -\overbrace{\sum_{n=1}^{N}}\ y_n \ln f_{xn} + 1 - y_n \ln 1 - f_{xn} \qquad 1.$$

其中N是AudioSet中的训练片段的数量。在训练中，通过使用梯度下降方法优化f（-）的参数，以最小化损失函数l。

B. 数据网

1） 传统残差网络（ResNets）：在音频分类方面，较深的CNN已被证明比较浅的CNN具有更好的性能[31]。非常深的传统卷积神经网络的一个挑战是，梯度不能从顶层正确地传播到底层[32]。为了解决这个问题，ResNets[32]在卷积层之间引入了快捷连接。这样，前向和后向信号可以直接从一层传播到任何其他层。快捷连接只引入了少量的额外参数和一点额外的计算复杂性。ResNet由几个块组成，其中每个块由两个卷积块组成。内核大小为3.3的层以及输入和输出之间的快捷连接。每个瓶颈块由三个卷积层组成，网络架构中有一个网络[39]，可以用来代替ResNet中的基本块[32]。

2） 调整ResNets进行AudioSet标记：我们将ResNet[32]调整为AudioSet标记，如下所示。首先，在log-mel频谱图上应用两个卷积层和一个下采样层，以减小输入log-mel谱图的大小。我们实现了三种不同深度的ResNet：一个22层的ResNet，有8个基本块；具有16个基本块的38层ResNet和具有16个残差块的54层ResNet。表II显示了适用于AudioSet标记的ResNet系统的架构。BasicB和BottleneckB分别是基本块和瓶颈块的缩写。

C. 移动网络

1） 传统移动网络：当系统在便携式设备上实现时，计算复杂性是一个重要问题。与CNN和ResNets相比，MobileNets旨在减少CNN中的参数数量和乘加操作，MobileNets基于深度可分离卷积[40]，通过将标准卷积分解为深度卷积和11点卷积[40]。

表三显示了AudioSet[40]的MobileNetV1和MobileNetV2[41]架构。V1Blocks和V2Blocks是MobileNet卷积块[40][41]，每个卷积块分别由两个和三个卷积层组成。

2) *Adapting MobileNets for AudioSet tagging* We adapt

D. 一维CNN

以前的音频标记系统基于log-mel频谱图，这是一种手工制作的功能。提高性能。一些研究人员提出构建直接对时域波形进行操作的一维CNN。例如，Dai等人[31]提出了用于声学场景分类的一维CNN，Lee等人[42]提出了一种一维CNN，后来Pons等人[15]将其用于音乐标签。

1） DaiNet:DaiNet[31]将长度为80、步幅为4的核应用于音频记录的输入波形。内核在训练过程中是可以学习的。首先，对第一卷积层应用最大值运算，该运算旨在使系统对输入信号的相移具有鲁棒性。然后，应用几个核大小为3、步幅为4的一维卷积块来提取高级特征。在UrbanSound8K分类中，每个卷积块中有四个卷积层的18层DaiNet取得了最佳结果[43][31]。

2） LeeNet：与在第一层应用大内核的DaiNet相反，LeeNet[42]在波形上应用了长度为3的小内核，以代替频谱图的STFT