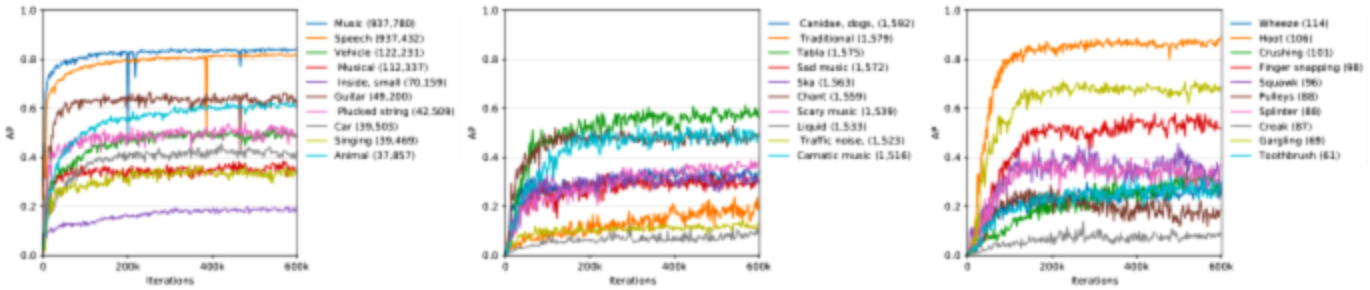

图3。使用CNN14系统对声音事件进行分级AP。括号内的数字表示训练片段的数量。左、中、右列显示了声音类的AP，训练片段的数量在AudioSet的训练集中排名第1至第10、第250至第260和第517至第527。

A. AudioSet数据集

AudioSet是一个大规模的音频数据集，包含527个声音类[1]。AudioSet中的音频片段是从YouTube视频中提取的。训练集由2063839个音频片段组成，包括22160个音频片段的平衡子集，其中每个声音类别至少有50个音频片段。评估集包含20371个音频片段。我们没有使用[1]提供的嵌入功能，而是在2018年12月通过[1]提供的链接下载了AudioSet的原始音频波形，并忽略了不再可下载的音频片段。我们成功下载了1934187个（94%）完整训练集的音频片段，包括20550个（93%）平衡训练集的视频片段。我们成功下载了18887个评估数据集的音频片段。如果音频片段短于10秒，我们会将其静音至10秒。考虑到YouTube上的大量音频片段是单声道的，采样率低，我们将所有音频片段转换为单声道，并将其重新采样为32 kHz。

对于基于log-mel谱图的CNN系统，STFT应用于大小为1024的汉明窗口[33]和跳数为320个样本的波形。这种配置导致每秒100帧。根据[33]，我们应用64个梅尔滤波器组来计算对数梅尔谱图。梅尔组的下限和上限截止频率分别设置为50 Hz和14 kHz，以消除低频噪声和混叠效应。我们使用torchlibrosa，librosa[46]函数的PyTorch实现，将log-mel光谱图提取构建到PANNs中。10秒音频片段的log-mel频谱图具有1001 64的形状。额外的一帧是在计算STFT时应用中心参数造成的。批量大小为32，使用学习率为0.o01的Adam[47]优化器进行训练。系统在一张Tesla -V100-PCI-32GB卡上进行训练。每个系统需要大约3天的时间从头开始训练600k次迭代。

B. 评价指标

平均精度（mAP）、平均曲线下面积（mAUC）和d-prime被用作AudioSet标记的官方评估指标[20][1]。平均精度（AP）是

表四
与以往方法的比较

	mAP	AUC	d-prime
随机猜测	0.005	0.500	0
谷歌CNN[1]	0.314	0.959	2.452
单级关注[16]	0.337	0.968	2.612
多层次关注[17]	0.360	0.970	2.660
功能级别关注度高[20]	0.369	0.969	2.640
好未来网[19]	0.362	0.965	2.554
DeepRes[48]	0.392	160	2.682
我们提出的CNN14	0.431	0.973	2.732

召回率和精确度曲线下的面积。AP不依赖于真阴性的数量，因为精确性和召回率都不依赖于真实阴性的数量。另一方面，AUC是假阳性率和真阳性率（召回率）下的面积，反映了真阴性的影响。d-prime[1]也用作度量，可以直接从AUC[1]计算。所有指标都是在单个类上计算的，然后在所有类上取平均值。这些指标也称为宏观指标。

C. AudioSet标记结果

I） 与先前方法的比较：表四显示了我们提出的CNN14系统与先前Au-dioSet标记系统的比较。我们选择CNN14作为基本模型来研究AudioSet标记的各种超参数配置，因为CNN14是一个结构简单的标准CNN，可以与之前的CNN系统进行比较[3][33]。作为基线，随机猜测分别达到0.005的mAP、0.500的AUC和0.000的d-prime。谷歌[1]发布的结果使用[13]中的嵌入特征进行训练，分别达到了0.314的mAP和0.959的AUC。单级注意力和多级注意力系统[16]、[17]实现了0.337和0.360的mAP，后来通过特征级注意力神经网络进行了改进，实现了0.369的mAP。Wang等人[19]研究了五种不同类型的注意力功能，并获得了0.362的mAP。上述所有系统都是基于AudioSet发布的嵌入功能构建的[1]。最近的DeepRes系统[48]是基于从YouTube下载的波形构建的，并实现了0.392的mAP。表四的底部行显示了我们提出的

*https://github.com/qiuqiangkong/torchlibrosa