IV、 数据处理

在本节中，我们将介绍AudioSet标记的数据处理，包括数据平衡和数据增强。数据平衡是一种用于在高度不平衡的数据集上训练神经网络的技术。数据增强是一种用于增强数据集的技术，以防止系统在训练过程中过度拟合。

A. 数据平衡

可用于训练的音频片段数量因声音类别而异。例如，有90多万个音频片段属于“演讲”和“音乐”类别。另一方面，只有几十个音频片段属于“牙刷”类别。不同声音类别的音频片段数量呈长尾分布。在训练过程中，训练数据以小批量输入到PANN中。如果没有数据平衡策略，音频片段将从AudioSet中均匀采样。因此，在训练过程中更有可能对具有更多训练片段的声音类（如“语音”）进行采样。在极端情况下，一个小批量中的所有数据可能属于同一个声音类。这将导致PANN过度适应具有更多训练剪辑的声音类，而不足适应具有更少训练剪辑的音频类。为了解决这个问题，我们设计了一种平衡的采样策略来训练PANN。也就是说，从所有声音类中大致相等地采样音频剪辑以构成一个小批。我们使用“近似”一词是因为音频剪辑可能包含多个标签。

B. 数据扩充

数据增强是防止系统过度拟合的有效方法。AudioSet中的一些声音类只包含少量（例如数百个）训练片段，这可能会限制PANN的性能。我们在训练过程中应用mixup[44]和SpecAugment[45]来增强数据。

1） Mixup:Mixup[44]是一种通过插值数据集中两个音频片段的输入和目标来增强数据集的方法。例如，我们将两个音频片段的输入分别表示为x1，2，其目标分别表示为y1 y2。然后，可以分别通过z=Ax1+（1A）2和y=Ay1+（2A）y2获得增强输入和目标，其中A是从贝塔分布中采样的[44]。默认情况下，我们对log-mel频谱图应用mixup。我们将在第VI-C4节中比较log-mel频谱图和时域波形上的混频增强性能。

2） SpecAugment：SpecAugment[45]被提出用于对语音数据进行增强以进行语音识别。SpecAugment频率掩蔽和时间掩蔽，应用频率掩蔽，使得f个连续的梅尔频率区间[fo，fo+f]被掩蔽，其中f从0到频率掩蔽参数f的均匀分布中选择，fo从[0，Ff]中选择，其中f是梅尔频率区间的数量[45]。每个log-mel频谱图中可以有多个频率掩模。频率掩模可以提高PANN对音频片段频率失真的鲁棒性[45]。时间掩蔽类似于频率掩蔽，但应用于时域。。

operates on the log-mel spectrogram of an audio clip using

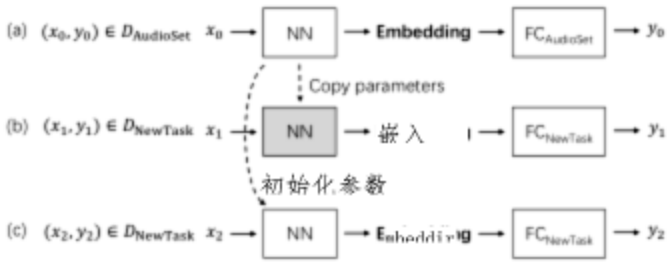

图2。（a） PANN使用AudioSet数据集进行预训练。（b） 对于新任务，PANN被用作特征提取器。基于提取的嵌入特征构建分类器。阴影矩形表示参数已冻结且未训练。（c） 对于新任务，神经网络的参数用PANN初始化。然后，对新任务的所有参数进行微调。

V. 转移到其他任务

为了研究PANNs的泛化能力，我们将PANNs转移到一系列音频模式识别任务中。之前关于音频迁移学习的研究[21][14][22][25][23]主要集中在音乐标记上，并且仅限于比AudioSet更小的数据集。首先，我们在图2（a）中演示了PANN的训练。这里，DAudioSet是AudioSet数据集，xo、yo分别是训练输入和目标。FCAudioset是AudioSet标签的全连接层。在这篇文章中，我们建议比较以下迁移学习策略。

1） 从头开始训练一个系统。所有参数都是随机初始化的。系统类似于PANN，除了最终的全连接层取决于任务相关的输出数量。该系统被用作与其他迁移学习系统进行比较的基线系统。

2） 使用PANN作为特征提取器。对于新任务
通过使用
潘。然后，嵌入特征被用作输入
分类器，如全连接神经网络
在训练新任务时，PANN的参数被冻结
没有受过训练。仅构建分类器的参数
对嵌入特征进行训练。图2（b）显示了这一点
策略，其中DNewTask是一个新的任务数据集，FCNewTask
是新任务的完全连接层。PANN用作
特征提取器。基于提取的嵌入构建分类器
丁特征。阴影矩形表示参数
它们是冷冻的，没有经过训练。。

（3） 微调PANN。PANN用于新任务，但最终的全连接层除外。所有参数都从PANN初始化，除了随机初始化的最终全连接层。所有参数都在DNewTask上进行了微调。图2（c）展示了PANN的微调。

VI、 实验

首先，我们评估了PANN在AudioSet标记上的性能。然后，将PANN转移到几个音频模式识别任务中，包括声学场景分类、通用音频标记、音乐分类和语音情感分类。