提取。LeeNet由几个一维演化层组成，每个层后面都有一个大小为2的下采样层。原始的LeeNet由11层组成。

3） 将一维CNN用于AudioSet标记：
我们修改了LeeNet，将其扩展到具有24层的更深层次的架构，用由两个卷积层组成的卷积块替换每个卷积层。为了进一步增加一维CNN的层数，我们提出了一种核大小为3的一维残差网络（Res1dNet）。我们用残差块替换LeeNet中的卷积块，其中每个残差块由两个核大小为3的卷积层组成。卷积块的第一和第二卷积层分别具有1和2的膨胀，以增加相应残差块的感受野。在每个残差块之后应用下采样。通过使用14个和24个残差块，我们分别获得了31层和51层的Res1dNet31和Res1dNet 51

III、 波形-卷积神经网络系统

以前的一维CNN系统[31][42][15]并没有优于以log-mel谱图作为输入训练的系统。以前的时域CNN系统[31][42]的一个特点是，它们不是为了捕获频率信息而设计的，因为一维CNN系统中没有频率轴，所以它们无法捕获具有不同音调偏移的声音事件的频率模式。

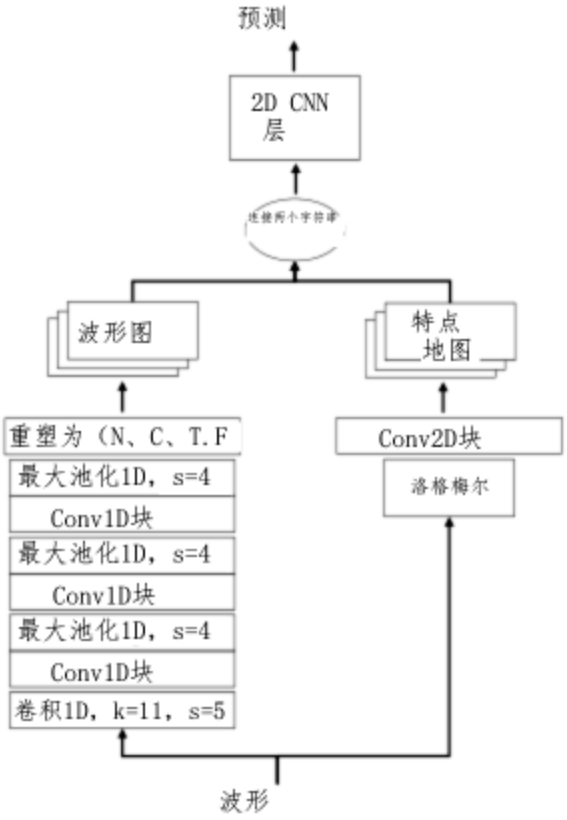

图1。Wavegram-Logmel CNN的架构

A. 波形-CNN系统

在本节中，我们提出了用于音频集标记的Wavegram CNN和Wavegram Logmel CNN架构。我们提出的Wavegram CNN是一种时域音频标记系统。我们提出的波形图是一种类似于log-mel频谱图的特征，但是使用神经网络学习的。波形图旨在学习傅里叶变换的修改后的时频表示。波形图具有时间轴和频率轴。频率模式对于音频模式识别很重要，例如，具有不同音调偏移的声音属于同一类。Wavegram旨在学习一维CNN系统中可能缺乏的频率信息。波图还可以通过从数据中学习一种新的时频变换来改进手工制作的log-mel谱图。然后，波形图可以代替对数梅尔谱图作为输入特征，从而形成我们的波形图-CNN系统。我们还将Wavegram和log-mel光谱图结合起来作为一个新特征，构建了Wavegram Logmel-CNN系统，如图1所示。

为了构建波形图，我们首先将一维CNN应用于时域波形。一维CNN从具有滤波器长度11和步长5的卷积层开始，以减小输入的大小。这会立即将输入长度减少5倍，以减少内存使用。接下来是三个卷积块，其中每个卷积块由两个卷积层组成，卷积层的膨胀分别为1和2，它们是设计用于增加卷积层的接收场。每个卷积块后面都有一个步幅为4的降采样层。通过使用步幅和降采样三次，我们将32kHz的音频记录降采样到每秒32000/5/4/4=100帧的特征。我们将一维CNN层的输出大小表示为TC，其中T是帧数，C是信道数。我们通过将C通道拆分为C/F组，将此输出重塑为T F C/F大小的张量，其中每组有F个频率区间。我们称这个张量为波图。Wavegram通过在每个C/F信道中引入F个频率仓来学习频率信息。我们在提取的Wavegram上应用第II-A节中描述的CNN14作为骨干架构，以便我们可以公平地比较基于Wavegram和log-mel频谱图的系统。二维CNN（如CNN14）可以捕获波形图上的时频不变模式，因为核在波形图中沿着时间和频率轴进行卷积。

B. Wavegram-Logmel-CNN

此外，我们可以将波形图和对数梅尔谱图组合成一个新的表示。通过这种方式，我们可以利用时域波形和对数梅尔谱图中的信息。组合是沿着通道维度进行的。Wavegram为音频标记提供了额外的信息，补充了log-mel频谱图。图1显示了Wavegram Logmel CNN的架构。