## 提取音频数据的梅尔频谱图

音频分析方法通常基于音频信号中提取的声学特征来实现对声音的紧凑且无冗余的表达，不仅有利于降低分析系统所需的输出存储和计算需求，更重要的是有助于机器学习算法对目标应用的实现。

利用librosa库来进行梅尔频谱系数的提取。本文取滤波器组个数=224，FFT窗口长度=1024，采样率为22050kHz，随后转换到对数刻度，得到对数梅尔频谱图。

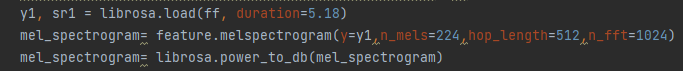


图 1 提取梅尔频谱图

求出对数梅尔频谱图的均值与方差，对每一个数据都进行的标准化。标准化后的数据如下图所示。

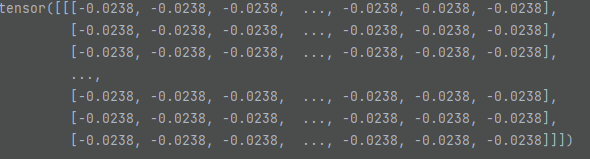


图 2 狗叫声对数梅尔频谱图数据标准化

将标准化后的部分数据可视化，得到频谱图如下所示。



图 3 狗叫声（a）与鸟叫声（b）的梅尔频谱图可视化

## 数据增强

对梅尔频谱图进行添加噪声，时间扭曲、时间掩蔽和频率掩蔽等数据增强操作，以增加数据量和避免模型的过拟合。

### 产生随机噪声

利用np.random()函数产生随机噪声，添加到原始数据中。

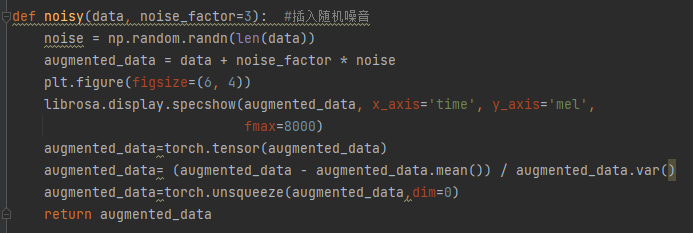


图 4 添加噪音函数



图 5 添加噪声前（a）后（b）对比图

### 时间扭曲

给定一个具有*τ*时间步长的对数梅尔频谱图，其视为时间轴为水平而频率轴为垂直的图像。在时间步长C:\Users\猪猪侠\AppData\Local\Temp\ksohtml11052\wps1.png内穿过图像中心的水平线上的随机点将被向左或向右扭曲，距离*W*从0到时间扭曲参数*W*沿该线的均匀分布中选择。

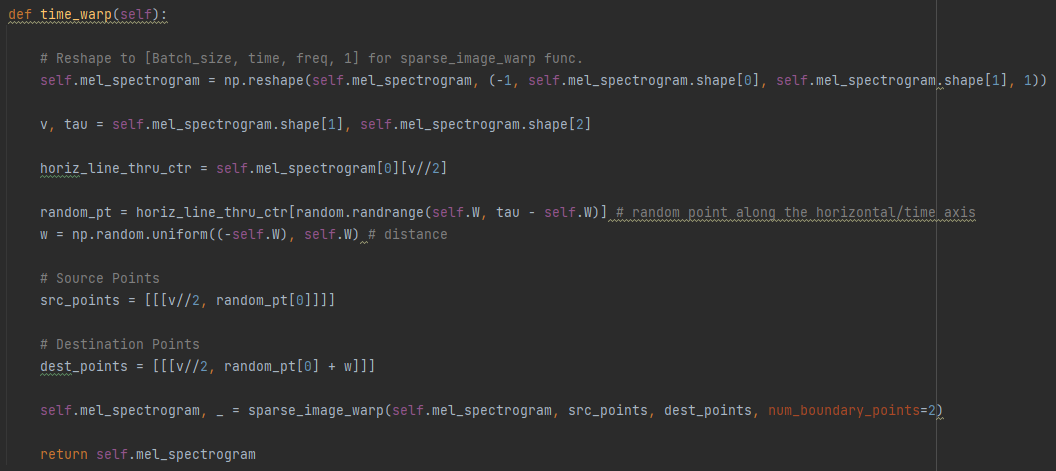


图 6 时间扭曲函数



图 7 时间扭曲前（a）后（b）对比图

### 时频掩蔽

时间掩蔽是指随机选择时间频率图中的一些时间步，并将这些时间步的所有频率分量都设置为0，以模拟掩蔽效应。而频率掩蔽则是随机选择时间频率图中的一些频率分量，并将这些频率分量的所有时间步都设置为0。

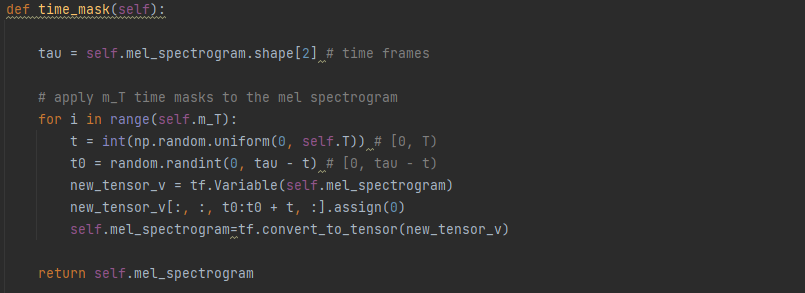


图 8 时间掩蔽函数

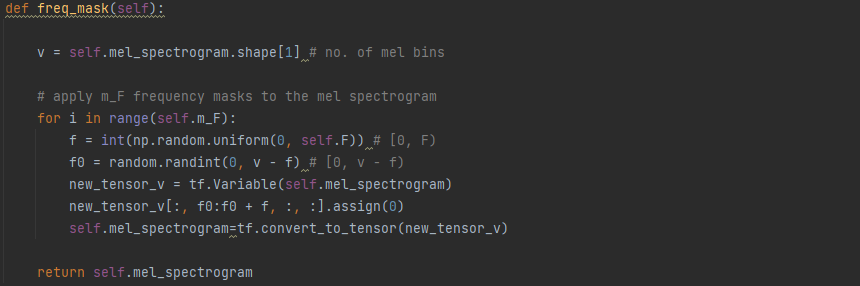


图 9 频率掩蔽函数



图 10 时间掩蔽（a）与频率掩蔽（b）

## Conformer模型的训练

模型的训练选取Adam优化函数，损失函数为CrossEntropyLoss，初始学习率为5e-5，每一轮使学习率衰退0.95，训练30个epochs。将音频数据转换为梅尔频谱图后整合为数据集，输入Conformer模型进行训练。

将224x224的梅尔频谱图数据输入模型，通过stem模块、第一个CNN模块和Transformer模块后得到的数据形状如图所示。

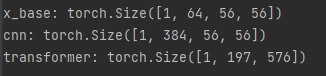


图 11 初始数据形状

在经过CNN分支和Transformer分支提取特征后，两条分支的数据形状并不一致，需要FCU模块进行分支特征的耦合。



图 12 数据形状差异

最后得到的训练损失以及精度结果如下图所示。

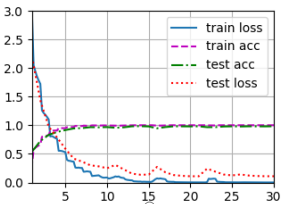


图 13 Conformer训练损失及精度曲线

## 消融实验

将在ESC-50数据集相同条件设置下训练好的ResNet50、Vit-base和本文的Conformer96模型上对模型输出的特征热力图作比较。对比结果如图 15所示，颜色越深的地方说明模型的关注度越高。Conformer可以很好的关注到梅尔频谱图的关键特征。

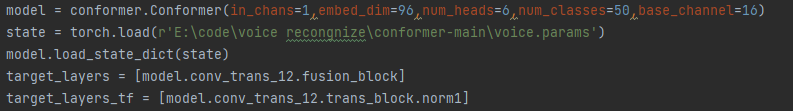


图 14 加载训练好的模型参数

图 15 特征热力图对比