视听信息系统导论大作业报告

漆耘含　　陶云松
2016011058　　2016013338

1　课程设计要求

给定若干乐器合奏视频以及独奏视频，同学可以利用各种方法设计出视觉信息和听觉信息的单独或联合表达特征，在视频和音频层次获取音源信息，实现乐器合奏视频的音源分离与定位。

2　原理及实现

本次大作业主要是参考《Learning to Separate Object Sounds by Watching Unlabeled Video》这篇论文而进行的设计。

本次大作业由两个部分构成：音源分离、图像特征提取。图像特征是通过densenet-161模型来提取的，通过将图像划分为左右两个部分，分别送进网络训练，然后取概率最大的乐器，作为一个标签，最终可以得到每一张图像的左、右两个标签，这个标签会在音源分离中用到。音源分离部分，先通过对solo部分的音频做NMF分解，再通过MIML网络，训练一个分类器，并从中训练得到每一种乐器的基，再通过指导NMF分解还原音频。下面将详细阐述这两个部分。

2.1　图像特征提取

图像特征是通过pytorch框架下的densenet-161模型来提取的，将图像送入已经训练好的模型进行分析，得到该图像对应的八种乐器的概率，取出概率最大值对应的乐器作为该图像的乐器标签。

为了分别获取图像左右部分乐器的标签，我们将每张图片分解成左、右两张图片，分别送入模型进行分析，得到图片左右两边的乐器标签，该标签将用于指导后续的音频分离。

2.2　音源分离

2.2.1　训练音频特征提取

本次大作业使用的训练集全部是solo的音频，音频是以采样率等于44100Hz进行采样的时间序列s(t)，要提取其音频特征，需要先对原始音频做短时傅里叶变换(STFT)，得到时频图$V_{F\times N}$，F是频率的个数，N是frames的个数，可以从时频图看到频率随时间的变化，短时傅里叶变化的原理在这里不做过多解释。得到时频图后再取其模值，进行非负矩阵分解（NMF），可以将V分解为两个矩阵的乘积：$V = W_{F\times M}H_{M\times N}$，M是提前设定的，W矩阵中的每一列都是V的一个基，H是相应的权重矩阵。对于每一个unlabeled的音频，都可以通过STFT和NMF提取到基矩阵W，这是作为MIML网络的输入。

一个音频的长度都有好几分钟，如果对整个音频做STFT和NMF，提取到的特征可能会不太理想，因此我们采用的方法是，先对整个音频做STFT，再将时频矩阵V在N这一维划分成若干个小矩阵，每一个小矩阵对应时域中的5s，这样分段进行特征提取，得到的基会比较好。

2.2.2　神经网络设计

因为得到的训练集是unlabeled的solo视频，为了将提取的基和乐器成功关联起来，这里采用MIML神经网络，即多目标多标签神经网络。

神经网络的input是从一个音频提取出来的基矩阵$W_{F\times M}$，然后通过一个孪生网络，即shared weights，孪生网络的目的是对音频的基降维，通过全连接层(FC)+batch norm(BN)+ReLU，学习音频的模式，然后将所有的输出组成一个特征矩阵$Y_{1024\times M}$。然后将特征矩阵Y通过一个1×1 Convolution-BN-ReLU模块，reshape成一个特征三维矩阵$R_{K\times L\times M}$，其中，K是每一种类别乐器下的子类别数量，比如乐器吉他有很多种类，比如电音吉他、木吉他等；L是需要预测的乐器种类，比如吉他、喇叭等。对特征三维矩阵在K这一维做Maxpooling，得到特征图$X_{L\times M}$，再对M这一维做Maxpooling可以得到乐器种类预测矩阵$D_{L\times 1}$，矩阵中最大值对应了预测的乐器种类。

训练集是没有进行标记的，因此为了训练一个分类器，需要图像特征提取，但图像识别准确率没有达到百分之百，训练的时候效果不好，因此为了完成本任务，训练集的标签采用人为标定的方法，使训练不会出错。在进行训练的时候，采取的损失函数是NLL指标。

设计这个神经网络，目的是提取出每一种乐器的特征基。提取特征基的方法，需要利用第一次Maxpooling之后的特征矩阵$X_{L\times M}$，L这一维代表的是乐器种类，M这一维代表的是基，特征矩阵X相当于是一个乐器-基关联矩阵，假设最后预测得到的是吉他，则X矩阵中吉他那一行的最大值对应的Mth基，即认为是吉他乐器的一个特征基。在训练结束后，对每一个预测

Algorithm 1 神经网络流程

```
for i in train epoch:
    for j in batch num:
        输入一个batch的基矩阵
        通过FC+BN+ReLU得到Y
        通过Convolution-BN-ReLU得到R
        对K所在维进行Maxpooling得到X
        对M所在维进行Maxpooling得到预测矩阵D
    计算loss，并后向传播更新参数
    计算测试集的准确率(accuracy)
计算所有训练集的准确度(accuracy)
for z in audio num:
    通过神经网络得到特征图X和预测结果i
    if i == label[i]:
        挑选出特征图中对应的基
保存提取到的每一种乐器的特征基
训练完成结束
```

正确的音频，都提取最可能的一个基，最后每一种乐器都可以得到一个特征基矩阵W_i，i = 0,1,...,L-1。这些基矩阵会在音源分解的时候用到。

2.2.3 音源分离

在音源分离部分，主要是利用guide-NMF分解。

先通过图像特征抓取模型得到待分离视频左边和右边的乐器标签，然后对待分离音频s(t)进行STFT分解得到矩阵V，类似地，对N这一维进行划分，对应到时域同样是5s。拿出经过神经网络训练得到的特征基（只拿出图像预测得到的两种乐器的基）拼合在一起，再进行guide-NMF分解，即已知STFT矩阵V，和基矩阵W，通过迭代得到权重矩阵H:

$$V_{F\times N_i} = W^q H^q = [W_1^q, W_2^q][H_1^q, H_2^q]^T$$

Algorithm 2 神经网络流程

```
加载每一种乐器的特征基
for i in audio num:
    对audio进行STFT并分段
    for j in audio slices num:
        进行guide-NMF
        计算并归一化得到V_j1，V_j2
    将每一段得到的V拼接分别得到V_1，V_2
    分别对V_1，V_2进行ISTFT得到分离音频
    保存音频
程序结束
```

然后可以相应地得到两个基对应的矩阵：$V_i^q = W_i^q H_i^q$，i = 1,2。再进行归一化操作：$V_i = \frac{V_i^q}{V_1^q+V_2^q} V_{F\times N_i}$，其中，$V_{F\times N_i}$包含了该段音频中所有的幅度和相位信息。最后将两种乐器对应的V_i按顺序拼接起来，分别进行逆短时傅里叶变换（ISTFT），得到分离的左边和右边对应的音频。

2.3 过程及结果展示

2.3.1 图像特征处理模型

我们利用crop函数将每张图片分割成左、右两部分，分别送入模型分析，最终得到左边和右边的乐器标签。我们不妨分别将得到的左边与右边的乐器标签记为A和B，则有以下两种情况。第一种为A≠B，即我们从图片的左边和右边提取了两种不同的乐器特征，进而可以直接将(A,B)作为后续音频分离的标签；第二种为A=B，即该图片对应一个独奏场景，这显然是不对的。产生这种结果的原因为左边或者右边的演奏者图片太大，拼接时该图片占据了整个图片的大部分，以至另一个演奏者图片太小而不好识别。这时，我们考察从左图和右图得到的乐器概率矢量，倘若A乐器在左图的概率大于右图，我们就认为A为左图的乐器标签，然后将右图继续分割为左、右两部分，将右部分再次送入网络，得到新的乐器标签C。倘若此时有A≠C，我们就将C作为原始图片右边的乐器标签，将（A,C）指导后续音频分离。倘若此时依旧有A=C，即从右图中识别出概率最高的乐器依旧是A，此时继续分割图像可能会导致图像过小而无法识别，因此我们取此时右图乐器概率矢量中概率第二高的乐器作为右图的标签D，将(A,D)指导后续音频分离。如果初始A乐器的在右图的概率大于左图则同理将左图进行再分割。

将图像送入模型进行分析时，每个文件的图片非常多，有几百甚至上千张，由于一个演奏视频中，相邻时间段演奏者的变化不会太大，因此将所有的图片均送入模型分析耗时太长，也没有必要。由于测试集是以每秒2.4帧提取的图像，我们决定等间隔选取200张图片，假设一个较长的演奏视频时长为10分钟，我们这样的选取每两张图片的时间间隔为3秒，基本可以认为3秒钟内演奏者动作变化不大，对于时长更短的视频误差会更小，因此这样选取图片的方法是可行的。

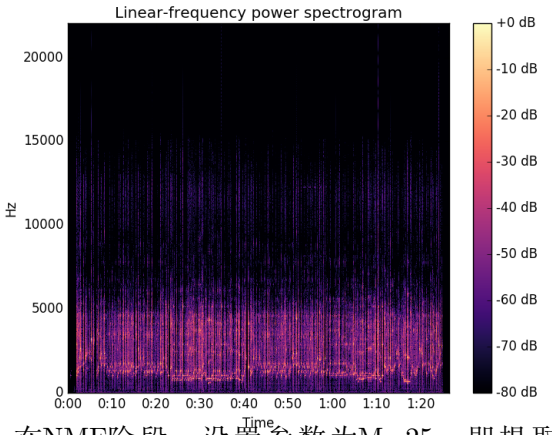
针对testset25测试集，用我们的方法得到的图像识别准确率为0.84，即25个视频，21个正确，4个错误。

2.3.2 音源分离模型

数据预处理对应的程序文件为STFT_NMF.py和nmf_data_process.py。

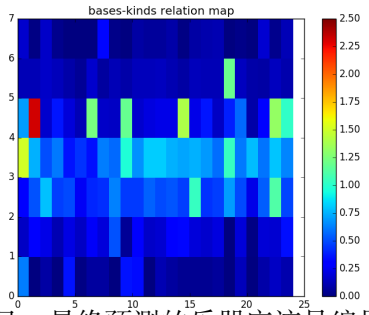
在solo音频预处理的时候，先进行STFT分解得到时频图V，然后分段进行NMF。其中，进行STFT的时候，是使用librosa.core.stft函数，参数设置为：n_fft = 4096，hop_length = 2048,即窗长为4096，重叠部分为窗长的一半，得到的V的大小为2049*Len，不同视频有不同的Len，2049为频率的点数。下图/data/dataset/audios/solo/xylophone/17.wav的时频

图:



在NMF阶段，设置参数为M=25，即提取出25条基。

神经网络的程序为Net.py。在神经网络训练过程中，使用的参数如下：L=8，M=25，K=4，base_length = 2049,batch_size = 64，train epoch=250，learning rate = 0.005。训练集有6976个基矩阵，测试集有1216个基矩阵。在训练过程中，计算了测试集准确度，达到0.88，在训练结束后计算了总体训练集的准确度，达到0.998。再根据$L \times M$特征图，提取基，下图为特征图:



由图可见，最终预测的乐器应该是编号5，并且应该选择第2个基作为该乐器的特征基。最终每一种乐器会提取到450-1200个特征基不等。

在音源分离阶段，同样是先对音频进行STFT变换，然后通过图像特征提取得到左和右的两个标签，再利用guide-NMF，固定V和W，其中W是对应乐器的特征基矩阵拼接起来的，然后通过迭代更新H，分别得到两种乐器的分离音频，具体方法见前面的小节，这里不再赘述。

针对test25测试集，我们方法得到的平均sdr为2.413dB，最好的能达到12dB，最差的会坏到-10dB，音频预测的准确率为0.96，即25个视频，24个正确，1个错误。

3 性能分析及问题

3.1 运行效率分析

在图像特征提取部分，对testset25进行特征提取，总共用时：4.5小时，双核CPU；在音源分离部分，神经网络训练（epoch=250）及挑选特征基运行时间为1.5小时，对testset25测试集进行音源分解运行时间为1.5小时，8核CPU。

最初时，我们并没有对图片进行抽样选取，直接将所有的图片送入模型分析，这样运行一次需要9个小时（双核CPU）。我们对图片进行等距离选取200张后，运行一次只需要4.5小时（双核CPU），且正确率保持不变，这样我们便提升了图像特征提取的运行效率。

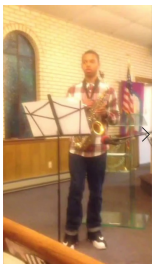
因为在音源分解部分用到了非负矩阵分解NMF，所以在运行效率上不是很高。在最开始的时候，我们的方法是对1分钟对应的音频进行分解，导致送进NMF分解的矩阵非常大，运行效率极低，并且效果也不太好，后来经过实验分析，发现在音频特征提取的时候，如果时间片划分得太长，提取出来的特征会不太好，而且运行效率也不高，因此我们将时间片划分为5s，运行效率大大提升，并且根据最后结果可知提取的特征基是比较好的。

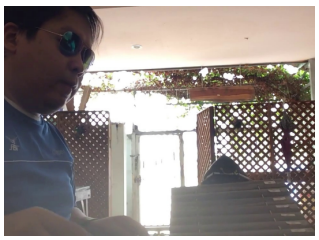
3.2 运行结果分析

3.2.1 图像处理结果

本次实验图像定位部分的准确率为0.84，即25组测试图片中有21组正确。错误的四组图片中，有三组中的半边图片来源于同一个吹萨克斯的演奏者（saxophone_1），我们的模型每次都将其识别为笛子；另一组错误的图片（flute_3_xylophone_2）中，我们的模型将木琴识别为了手风琴。

saxophone_1对应的一张典型图片如下图左侧所示。可见整张图片颜色偏黄，与萨克斯颜色接近，因此较难识别出该乐器，同时图片中存在许多杆状物，这也是将其识别为笛子的原因。



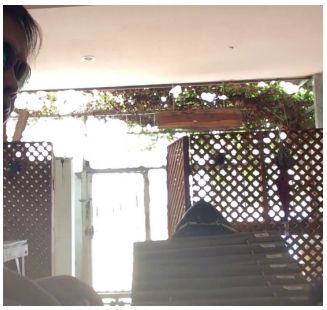


我们截取了10张该演奏者的图片，进行独奏识别测试，得到的乐器概率矢量为[0.02921172，0.00188907，0.00941814，0.17900135，0.3473081，0.00477731，0.18529569，0.03099669]，按照我们对乐器标签的排序['accordion'，'acoustic_guitar'，'cello'，'trumpet'，'flute'，'xylophone'，'saxophone'，'violin']，判断笛子为乐器标签的概率最大，因此这张图片即使在独奏测试时也无法将其识别。

同样的，xylophone_2对应的一张典型图片如上图右侧所示。该图中的木琴与手风琴样子的确比较接近。

我们同样截取了10张该演奏者的图片，进行独奏识别测试，得到的乐器概率矢量为[1.08197640e-01, 3.18867156e-04, 7.07219544e-05, 2.54932665e-03, 2.31028754e-03, 3.80916573e-01, 4.83326587e-03, 1.25244661e-03]，按照顺序可知，木琴的概率最大，独奏识别结果是正确的。

因此我们进一步分析出错的原因，我们将图片截取为二重奏时送入模型分析的大小，即原二重奏图的一半，图片如下图所示。此时进行独奏识别测试得到的乐器概率矢量为[7.99859546e-01, 4.82431392e-04, 9.82426964e-05, 2.15453774e-03, 7.24989075e-04, 3.96170468e-02, 3.76080798e-03, 3.80475882e-04]，这次手风琴的概率最大，因此我们得知二重奏两张图片大小的不均匀性对结果也有很大的影响。



总的来说，图像的识别准确率受到以下两个因素影响：一是该图中乐器在背景中的区分度以及图片中的干扰因素，如果乐器颜色与整体背景颜色相差不大或者图片中存在形状类似其他乐器的物品，识别效果便可能受到影响；二是合成的两张图片拼接形式，倘若一张图片占据了整张图片的大部分导致分割图片时进行了不恰当的分割，识别效果也会受到影响。但是我们在没有任何信息的情况下确实无法得知应该如何最优地按比例切割图像，这也是方案中需要提高的部分。

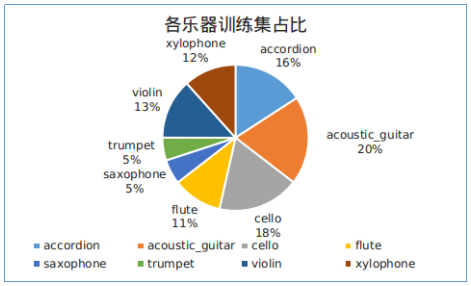
3.2.2 音源分离结果

本次实验最终的sdr为2.413dB。在最开始的时候，sdr只有-10dB，通过分析，大概确定了两点问题。第一点问题是音源特征提取。像上一小节所讲，在音频预处理的时候，先做STFT，再划分时间片进行NMF，这里的时间片是1分钟，时间片过长会造成提取到的乐器特征基不太好，并且运行效率也很低，因此我们将时间片划分为20s，10s和5s进行实验，效果都是不错的，最后采取的是5s的时间片长度。

第二点问题是在神经网络训练结束之后找特征基。因为基的选择与后面音源分解的效果大大相关，我们的方法在选择基的时候，发现了一些问题。第一个是基的对应问题，挑选特征基是从特征图中挑选的，因此特征图M（基）的顺序应该和最开始送进网络的M（基）顺序一致，如果顺序错乱，则添加的是错误的特征基，为了保证顺序是一致的，我们的方法是网络中M对应的那一维保持不变，这样就保证了基的顺序是一致的。第二个是在所有的训练集中，每一种乐器都会挑出450-1200个基不等，音源分离的时候只用到其中的400个基，为了保证效果，即用比较好的基，我们在从特征图挑选基的时候，把对应特征图中的对应数值也记下来，分乐器种类对数值进行排序，取最大的400个数值对应的特征基，这样分离的效果会更好。

最终sdr结果为2.413dB，通过对每一个音频的sdr进行输出，可以看到sdr结果极差比较大，即效果好的可以到10dB以上，比如：吉他，小提琴；但效果不好的也会到-10dB，比如：萨克斯，喇叭。

其中，喇叭的分解效果不太理想，经过分析，得到如下结论，即提取出来的喇叭的特征基数量为420，而吉他、小提琴等的特征基数量有1200个左右，喇叭的可供选择的基比较少，其中可能有不太好的特征基。在神经网络训练的时候，送进去的训练集各个乐器所占比如下所示：



可以看到，喇叭和萨克斯的占比是比较少的，吉他的占比是喇叭和萨克斯的4倍，因此在训练神经网络的时候，这两者得到的训练相对不足。有一个解决思路是减小占比大的乐器数量，保证送进去训练的乐器占比差不多相等，但因为受最小占比的限制，会导致提取出来的各个乐器的特征基矩阵比较小，效果反而会更差。因此，如果能增多萨克斯和喇叭的训练集，效果应该会好不少，但因为训练集是给定的，因此无法实际验证。

4 课程设计总结

本次课程实验，使用到了神经网络，我们小组之前都没有接触过，因此都是从零开始，通过查询大量资料、实践，最终成功完成了本次大作业，让我们对神经网络的基本结构、训练过程、优化等有更深的认识，并且对神经网络编写的细节也有所了解，比如对训练集和测试集的划分，在刚开始的时候就应该分离出固定的一部分作为测试数据。同时，本次大作业还用到了STFT、NMF以及densenet-161模型，在调试过程中了解了STFT窗长和步长对变换的影响，NMF分解的运行效率，以及densenet模型中mean和std的生成方式。

本次大作业，实现了从0到1的突破，并且最终结果也还不错，非常感谢给予我们小组帮助的同学，也感谢助教对本次大作业的耐心答疑和批改!

5 分工

漆耘含：神经网络部分、音源分离部分

陶云松：图像特征提取部分