噪声干扰下的音视频匹配设计报告

小组成员及分工情况：

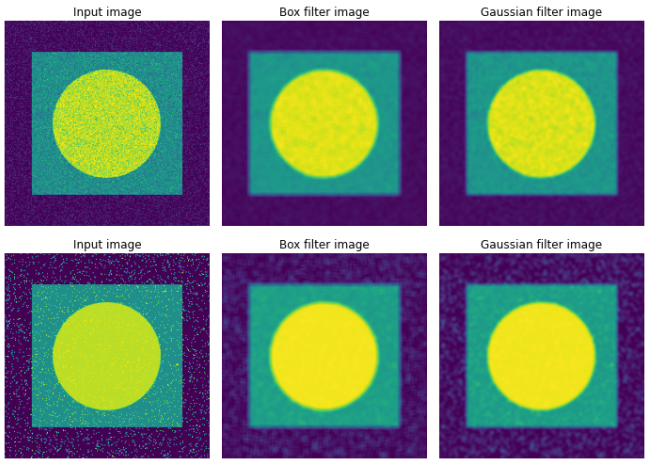
- 薛皓阳，无01，2020010647：完成任务一、撰写报告、测试特征；
- 张济达，无01， 2020010630：完成任务二、任务三神经网络设计与训练、撰写报告、检查结果；
- 刘畅，无01，2020010626：任务三降噪、提取特征、测试特征。

文件清单：

- `part1.ipynb`：任务一代码；
- `part2.ipynb`： 任务二代码；
- `Vfeat_extracter.py`：任务三（下省略）视频特征提取；
- `Vfeat_pca.py`：视频特征降维；
- `Afeat_ectracter.py`：音频特征提取；
- `v_denoise_feat.py`：视频降噪 + 特征提取；
- `a_denoise_feat.py` ： 音频降噪 + 特征提取；
- `similarity_test.ipynb`：检验输入特征的相关度；
- `denoise_test.ipynb`：检验音频去噪的效果；
- `ans_check.ipynb`：抽样相似度矩阵结果，人工检查匹配是否合理；
- `dataset.py`：数据加载；
- `models.py`：网络模型；
- `train.py`：训练网络；
- `test.py`：测试网络；
- `train_test.py`： 同时实现训练、测试网络得到批量测试结果。

任务一：基于滤波器的图像噪声处理

我们根据指导书实现了均值滤波器、高斯滤波器，并将两种图像去噪方法分别应用在两张含有不同种类噪声的图片上，进行对比分析。



图像去噪比较

按照滤波器不同比较

由两种方法得到的两种滤波结果可以看到，对于两种图片，均值滤波器和高斯滤波器结果相近

按照图片不同比较

将图中从外向内三个区域编号为1、2、3，则：

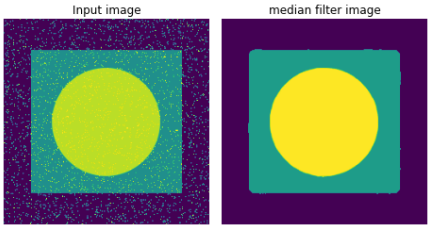
- 图一中区域1去噪效果较好，区域2一般，区域3较差；
- 图二中区域1去噪效果较差，区域2一般，区域3较好

分析

- 图一噪声幅度较小，区域颜色与噪声颜色相近程度为区域1 > 区域2 > 区域3，所以在区域1滤波效果较好，区域3滤波效果较差。
- 图二噪声幅度较大，区域颜色与噪声颜色相近程度为区域3 > 区域2 > 区域1，所以在区域3滤波效果较好，区域1滤波效果较差。

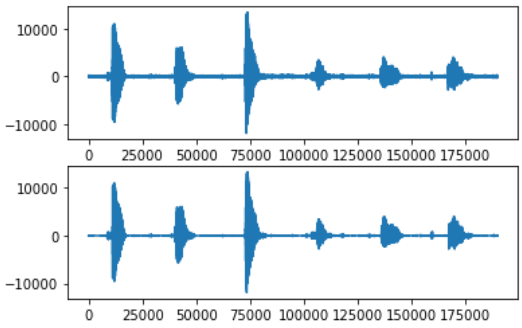
中值滤波结果分析

针对图二，该图所带噪声为比较明显的椒盐噪声，使用中值滤波可以有效消除。



任务二：基于谱减法的音频噪声处理

我们根据指导书实现了使用谱减法对带噪声音频的处理，基本流程是对输入语音信号$X[n]$加窗做分帧得到$X_i[n]$，做fft后取模方$|X_i[k]|^2$后与基于音频前5帧只含有噪声的先验假设下估计出的噪声频谱做差得到去噪后的频谱。（这里分纯净信号功率与噪声功率大小两种情况讨论，保证结果为正）将谱减后的结果开方得到干净频谱的幅度，相位则与$X_i[k]$保持一致。得到结果波形图如下所示：



可以观察到处理后（下）相较于处理前（上）在保持音频原有波形的基础上减弱了噪声波形，经过实际播放验证，音频降噪成功。

任务三：音视频匹配

基本原理

视频特征

- 对视频每1秒提取1帧图像，每个视频共提取10帧图像，每帧图像均送入特征提取网络（ResNet101），得到一个2048维(1×2048)的特征数据，则一个视频得到(10×2048)的特征矩阵
- 对前述得到的特征提取512维主成分分析（PCA），每视频最终产生(10×512)的特征。

音频特征

每条音频划分出10帧，每帧通过特征提取网络（Vggish）生成128维特征，最终每条音频产生(10×128)的特征。

网络结构

我们延续样例代码网络结构框架，对于输入的特征`Afeat`与`Vfeat`，先利用LSTM网络`AfeatRNN`对`Afeat`进行具有时域相关性编码，随后将特征分别送入匹配网络`Matching`，用于提取进行音视频匹配的特征，最后将两路数据送入匹配相似度网络`Prob`计算相似度。在样例代码基础上主要进行了如下改动：

1. 为避免LSTM网络的过拟合，不仅将时域相关性编码后的特征用于匹配网络的输入，也增加归一化后的直通分类`Afeat`作为输入；
2. 提高了`Matching`网络的输出维度与层数，并设置了`dropout`。

子任务一：清晰音视频匹配：

数据集划分

使用Train中Clean的音视频的95%作为训练集，在训练完成后马上用剩下5%用来测试（或验证），随机选取50对音视频计算Top1和Top5准确率。

训练 - 测试结果

经过100轮训练，得到：

- 在50条数据的测试集上Top1准确率为: 0.26892。
- 在50条数据的测试集上Top5准确率为: 0.77008。

分析

比对Top1匹配结果的音视频，可以发现清晰音视频匹配的结果基本正确，虽然Top1准确率只有不到27%，但由于数据集本身的音视频（尤其是音频）存在及格明显的分类，而类内相似度比较高，导致作为人类也无法准确判断语音视频是否匹配，只能准确地判断音视频是否同属于一类。

子任务二：带噪音视频匹配：

数据集划分

考虑到Train中未给出带噪声的音视频，我们使用Train中清晰的音视频生成了带噪声的音视频特征，并对特征划分训练集、数据集。其中随机选取50对音视频作为测试集，剩余的作为训练集。

数据集降噪

视频：考虑到视频特征提取时仅使用视频的10帧图像，因此直接在提取特征之前、截取视频之后对图像降噪即可，降噪使用`OpenCV`自带的方法`fastNlMeansDenoisingColored()`。

音频：最初尝试用Task2谱减法进行降噪，随后认为效果不佳，改用库`noisereduce`进行降噪。

训练 - 测试艰辛历程与结果

完成子任务二过程中有如下变量：网络参数；训练/测试降噪or不降噪；音频降噪方法。一一对其组合进行尝试与探索，并通`ans_check.ipynb`进行抽样检查，历程如下（其中测试集视频都经过降噪处理，音频因为更难降噪所以采用了不同的方法）：

1. 采用子任务一（即干净训练集得到）的网络参数，测试使用带噪提取的特征
    - 检查结果为音视频的类别都很不一致；
2. 采用子任务一的网络参数，测试使用task2谱减法降噪后音频提取的特征
    - 检查结果为音视频的类别都很不一致；
3. 采用子任务一的网络参数，测试使用`noisereduce`库降噪后音频提取的特征
    - 检查结果为音视频的类别都很不一致；
4. 训练和测试集同时task2谱减法降噪
    - 训练top5准确率仅20%；
    - 检查结果为音视频的类别都很不一致；
5. 训练和测试集同时`noisereduce`库降噪
    - 训练top5准确率仅20%；
    - 检查结果为音视频的类别都很不一致；
6. 心态崩了不对音频去噪，对训练音频加噪声提特征训练，用带噪特征测试
    - 训练top5准确率60%；
    - 检查结果有显著提升，不过还远不及子任务一结果。

分析与结论

- 对比1和子任务一的结果，说明子任务一训练的网络对噪声不鲁棒，若想更好地完成子任务二需要额外的训练；对比4，5和子任务一的训练结果，说明对音频去噪后提取的特征并不自洽，即去噪效果不好。它的原因如下：带噪音频的噪声过大，且前五帧并不一定是纯噪声，因而去噪后的音频距离干净的音频甚至比带噪音频还远（特征相关性检测表明）；以上也解释了2，3。子任务一中喂给网络的音频特征与测试中的去噪音频相差很多，结果自然不好；
- 子任务二相比子任务一难度显著提高，主要体现在音频去噪上。首先，音频去噪本身就难以非常准确、鲁棒地做到，尤其是当噪声很大的时候；其次，经过人工判别，受噪声影响不大的音频如赛车比赛、直升机、演讲等的匹配准确率较高，而一些本来就几乎无声的视频，如宠物则很难匹配。可以看出，音频之间的信噪比差异很大，而对于那些无声视频的信噪比更是趋近于零，可以想见，这样的匹配任务难度极大。综上，我们选取带噪音频用于训练，因为在对信噪比较低音频去噪之后得到的信噪比更低，只能单方面寄希望于网络能够自己学习识别噪声。问题的突破口，可能在神经网络的设计上，而这在我们的能力之外。可以说，在几乎没有受到课程给予机器学习知识指导的背景下，我们做了能想到的所有尝试。