《视听信息系统导论》课程大作业：

噪声干扰下的音视频匹配

2022 年 11 月 12 日

一． 问题背景

视频和音频是我们日常生活中接触最多的两种媒体形式。当我们听到一段优美的音乐时，我们往往会在脑海中绘制出美丽的画面，反之当我们看到某个场景时，也往往能想象出场景中可能出现的声音。这是因为我们人类具有跨通道知觉的能力，能够根据一种感觉特征确定另一感觉通道所熟悉的刺激物或形式。本次课程大作业给定若干对无声视频与音频文件，希望同学们使用机器学习方法，提取其视觉、听觉特征，并设计视听信息之间相似性的度量方法，评估无声视频与音频数据的匹配度，从而找回每个无声视频所对应的原始音频文件。除此以外，真实环境中采集的音视频信号往往存在噪声污染，在有噪声干扰的条件下如何实现准确的音视频匹配也是本次大作业需要研究的内容。

二． 数据集介绍

本次大作业提供 3339 组训练数据用于训练模型，另外分别提供 804 段**打乱的**视频数据和音频数据用作测试。每组训练数据包含**成对的视频、音频文件以及预先提取好的特征**，相同文件名前缀表示视频和音频来自同一段原始视频。数据可从清华云盘下载：https://cloud.tsinghua.edu.cn/d/8ff6b6deb394422e873d/，密码：std2022-2023。本次大作业的数据集含 2 个文件夹，分别为训练数据集 Train，测试集 Test，每个文件夹的内容如下：

- Train 文件夹：
  - video 文件夹：mp4 格式的无声视频文件，命名方式为%04d.mp4。
  - vfeat 文件夹：npy 格式的无声视频特征文件，命名方式为%04d.npy，每个文件内包含 10*512 维的特征。
  - audio 文件夹：wav 格式的音频文件，命名方式为%04d.wav。
  - afeat 文件夹：npy 格式的无声音频特征文件，命名方式为%04d.npy，每个文件内包含 10*128 维的特征

  上述文件夹中**相同前缀**的文件表示来自**同一原始视频**。
- Test 文件夹：
  - Clean 文件夹：无噪声干扰数据
    - video 文件夹：mp4 格式的无声视频文件，命名方式为%04d.mp4。
    - vfeat 文件夹：npy 格式的无声视频特征文件，命名方式为%04d.npy，每个文件内包含 10*512 维的特征。

    上述两个文件夹中**相同前缀**的文件表示来自**同一无声视频**。
    - audio 文件夹：wav 格式的音频文件，命名方式为%04d.wav。
    - afeat 文件夹：npy 格式的无声音频特征文件，命名方式为%04d.npy，每个文件内包含 10*128 维的特征。

    上述两个文件夹中**相同前缀**的文件表示来自**同一音频**。
  - Noise 文件夹：有噪声干扰数据
    - video 文件夹：mp4 格式的无声视频文件，命名方式为%04d.mp4。
    - vfeat 文件夹：npy 格式的无声视频特征文件，命名方式为%04d.npy，每个文件内包含 10*512 维的特征。

上述两个文件夹中**相同前缀**的文件表示来自**同一无声视频**。

- audio 文件夹：wav 格式的音频文件，命名方式为%04d.wav。
- afeat 文件夹：npy 格式的无声音频特征文件，命名方式为%04d.npy，每个文件内包含 10*128 维的特征。

上述两个文件夹中**相同前缀**的文件表示来自**同一音频**。

注意：测试数据中视频数据和音频数据之间并不匹配！要求同学们计算视、音频之间的匹配矩阵。

另外，本次大作业提供简短的特征提取代码，同学们可以根据自己需要决定是否使用提供的特征文件及特征提取代码。

三． 任务描述

1. 任务一：基于滤波器的图像噪声处理

日常拍摄的图像会由于环境设备等问题产生一系列的噪声，如果从噪声图像中恢复原有信息便成为了一个急需解决的问题。**这里给定含有一定噪声的图像，同学们将实现不同滤波器及其在图像去噪上的应用**。题目以 jupyter 的形式给出，代码内有详细描述和提示，同学们只需按要求完成相应模块的功能实现即可。建议使用 python3.6 及以上版本，所需 module 均为常规 module，可正常安装。任务所需数据均在 jupyter 文件同一路径下，直接可调用。

提交时注意提交 jupyter 完整执行结果，包含代码和输出结果。

2. 任务二：基于谱减法的音频噪声处理

音频同视频一样，由于环境因素和采集设备的缺陷，其也会携带一系列的噪声。与任务一类似，**这里给定含有噪声的一段音频，同学们将使用谱减法尽可能的将音频噪声降低**。题目同样以 jupyter 的形式给出，代码内有详细描述和提示，同学们只需按要求完成相应模块的功能实现即可。任务所需数据均在 jupyter 文件同一路径下，直接可调用。

提交时注意提交 jupyter 完整执行结果，包含代码和输出结果。

3. 任务三：音视频匹配

任务三要求同学们对视频和音频进行匹配，要求对于 N 段无声视频和 N 段音频进行相似度计算，返回$N \times N$的相似度矩阵 S，S_{ij}表示第 i 段无声视频和第 j 段音频的相似度。对于 S 的每一行，值最大的 k 个相似度对应的列为该行对应的无声视频的 top k 音频匹配结果。

任务提供了一份训练代码可供参考，提交时注意提交完整的模型定义代码、训练代码、模型文件，以及对 Clean 和 Noise 两类测试数据的相似度矩阵，按照 clean.npy 和 noise.npy 的格式保存。

四． 作业要求

1. 设计报告：

每小组提交一份设计报告，报告篇幅不得超过 4 页 A4 纸，报告应至少包含以下内容：

- 小组成员名单及分工情况：小组成员评分可能会因分工及完成情况产生差异。
- 提交文件清单。
- 工作开展及研究情况：应至少包含原理、实现方法、结果展示、结果分析、问题与不足，也可以包含其他任何对于解决问题有益的思考和讨论。

2. 提交清单：

每小组提交一份以“提交同学学号_提交同学姓名.zip/rar”命名的压缩文件，压缩文件内至少包含：

- 设计报告（.pdf/docx/doc）。
- 环境说明文件（.txt），包含主要依赖库的版本。

- 任务一、二的 jupyter 文件。
- 任务三完整的模型定义代码、训练代码、模型文件等，以及对 Clean 和 Noise 两类测试数据的相似度矩阵，按照 clean.npy 和 noise.npy 的格式保存。

五. 评分标准

1. 评价指标：

本次大作业的任务三使用 top k 匹配成功率作为评价指标。给定N × N的相似度矩阵 S，以及 ground truth 的匹配结果 GT, 匹配成功率定义为：

$$metric_k = \frac{\sum_{i=1}^{N} I(S_i, GT_i)}{N}$$

$$I(S_i, GT_i) = \begin{cases} 1, & GT_i \in argtopk(S_i) \\ 0, & else \end{cases}$$

2. 具体评分标准

本次大作业满分 100 分。满分 100 分由报告和结果两部分组成，其中报告占 30 分，结果占 70 分。根据三中的任务描述，结果分由两部分组成：任务一、二的音、视频去噪任务（30 分），任务三的音视频匹配任务（40 分），其中任务一、二正确填写 jupyter 代码并运行得到正确结果即可得满分，任务三根据 Clean 和 Noise 两项 top k 匹配成功率给分。

如有设计文件延期提交，设计报告、程序实现中存在抄袭行为等，将根据情节程度，扣除课程设计的部分或全部分数。