《视听信息系统导论》第三次大作业：

基于音频和图像序列的物体撞击匹配

2020 年 11 月 26 日

一. 问题背景：

近年来，计算机视觉和听觉技术在许多任务上都取得了巨大的成功，比如图像分类、物体检测，语音识别等。对于单个个体的建模往往是比较容易的，但对物体之间交互产生的视听信息进行建模则更为困难。在本次作业中，我们提供以下情境，如图 1 所示，物体放置于一个金属托盘中，托盘朝任意方向倾斜，物体与托盘壁发生碰撞并产生声音。在托盘壁四条边的中心各放置一个麦克风记录音频，摄像机在托盘的上方垂直向下捕获 RGB 图像。给定不同物体、不同运动状态的碰撞段落（每个段落包含以碰撞时刻为中心，持续时间为 4 秒的音频，以及对应的 RGB 图像序列），无论是音频还是图像序列中都蕴含着物体及其运动状态的特征，因此可以通过对上述特征进行合理的建模和学习，实现对音频和图像序列的匹配。请参考课程内容，根据二和三中的数据集和任务描述，对打散后的属于不同碰撞段落的音频和图像序列进行匹配。

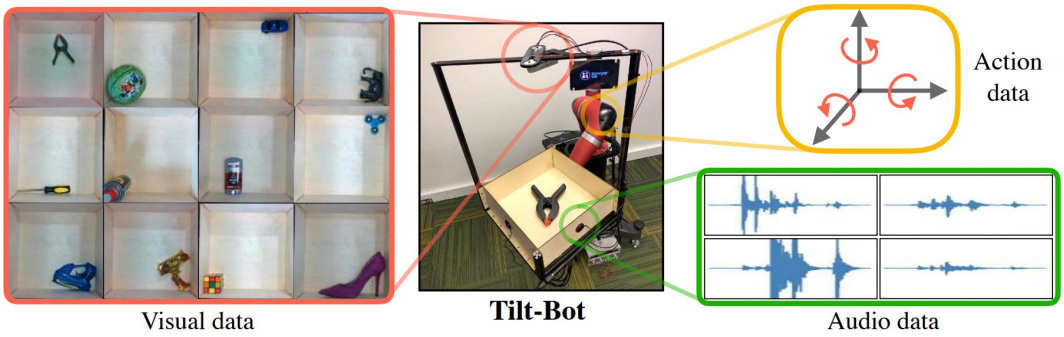

图 1

二. 数据集介绍：

本次大作业的数据集中包含 10 种不同的物体的撞击音频，每段音频截取自撞击发生时刻的前后各 2 秒内，四个通道的音频分别来自于托盘四条边中央位置的麦克风，每个通道音频的采样频率均为 44.1kHz。数据集还提供了与音频对应的 RGB 图像序列，每幅图像的尺寸为 480*640。为了方便同学们获得物体位置的信息，我们提供了与 RGB 图像序列对应的物体掩膜序列。物体掩膜序列通过对 RGB 序列进行简单的背景减除获得，为了排除托盘外背景对背景减除的影响，首先对 RGB 图像进行裁剪，左右各裁去 100 个像素，上下各裁去 20 个像素，得到的掩膜尺寸为 440*440。同学们可以自行决定是否使用其他方法提取更高质量的掩膜。训练数据和测试数据链接：

https://cloud.tsinghua.edu.cn/d/c1ba88cbcbbc4b0e9d5b/。密码：stdhomework3

由于清华云盘对单个文件大小的限制，下载下来的数据包含多个压缩文件，需要同学们自行解压。

训练数据组织方式如下：

```
./dataset/train                                  根目录
|   ./stanley_screwdriver                        物体名称
|   |   ./0                                      撞击片段
|   |   |   audio_data.pkl                       音频文件
|   |   |   ./rgb                                RGB 图像序列
|   |   |   ./mask                               掩膜序列
|   |   ./1
|   |   ⋮
|   ./green_basketball
|   ⋮
```

三. 任务描述：

1. 任务一：音频分类

不同物体的物理性质不同，因此撞击的音频也具有不同的特点。**任务一要求仅利用物体撞击的四通道音频信息对物体的种类进行分类，**要求对每个音频文件估计分类物体的编号。测试数据的组织方式如下：

```
./dataset/task1/test                             根目录
|   audio_0000.pkl                               音频文件
|   audio_0001.pkl
|   ⋮
```

为了方便测试，需要同学们参考 test.py 中定义的接口，在函数 **test_task1** 内部添加测试代码。test_task1 函数要求以测试数据的路径为输入，输出音频分类的结果，结果以字典的方式保存，比如：{‘audio_0000.pkl’：2, ‘audio_0001.pkl’：5, …}，以音频文件名为键，值为对应的类别序号（详见 test_task1 函数中的注释）。

2. 任务二：完全匹配

撞击音频不仅包含有关物体种类的信息，还包含物体撞击方向、位置等信息。考虑若干对匹配的音频和视频，将对应关系打乱之后，任务二要求同学们通过分析音频和视频的特征来恢复二者的对应关系。**任务二中待匹配的音频和视频数量相同，且音频和视频之间存在一一对应关系**。关于匹配问题，可以参考匈牙利算法、KM 算法等经典算法。任务二提供十组线下测试数据，另外助教还保留了十组线上测试数据，线上、线下数据具有相同的格式。任务二的测试数据如下组织：

```
./dataset/task2/test                             根目录
|   ./0                                          第 0 组测试数据
|   |   ./video_0000                             0 号视频
|   |   |   ./rgb                                图像序列
|   |   |   ./mask                               掩膜序列
|   |   ./video_0001
|   |   ⋮
|   |   audio_0000.pkl                           0 号音频文件
|   |   audio_0001.pkl
|   |   ⋮
|   ./1
|   ⋮
```

test.py 文件里的 **test_task2** 函数定义了任务二的输入输出接口，请同学们在函数**内部**添加代码。test_task2 函数要求以一组测试数据的路径为输入，输出该组测试数据的结果，结果以字典的方式保存，比如：{‘audio_0000.pkl’： 12, ‘audio_0001.pkl’：23, …}，以音频文件名为键，值为对应的视频序号(从 0 开始)。

3. 任务三：不完全匹配

任务三与任务二的目标相同，同样是对音频和视频进行匹配。**区别在于任务三不保证待匹配音频和视频一一对应：存在不确定数量的音频没有与之对应的视频，也存在不确定数量的视频没有与之对应的音频。**任务三的测试数据的根目录为./dataset/task3/test，文件结构与任务二相同。test.py 文件里的 **test_task3** 函数定义了任务三的输入输出接口。test_task3 函数要求以一组测试数据的路径为输入，输出该组测试数据的结果，结果以字典的方式保存，比如：{‘audio_0000.pkl’：12, ‘audio_0001.pkl’：-1, …}，以音频文件名为键，值为对应的视频序号(从 0 开始)，如果一段音频没有对应的视频，则值为-1。

同学们须严格按照函数 **test_task1、test_task2、test_task3** 定义好的**输入、输出格式**来组织代码，**不得改写输入、输出结构**。

四. 作业要求：

1. 设计报告：

本次作业分小组完成，每组成员不得超过 3 人。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（例如：torch==1.2.0）。
- 补充后的 test.py，包含函数 **test_task1、test_task2、test_task3**。
- 其他代码文件和依赖库文件。

3. 编程语言：

考虑到本次作业可能需要使用深度学习模型，因此建议使用 python。如果使用其他编程语言，**请按照三中要求（仿照 test.py 的形式）实现测试接口，并详细说明接口的调用方法。**

五. 评分标准：

1. 评价指标：

本次作业使用音频分类/匹配的准确率作为评价指标。对于分类任务，准确率为正确分类的音频占全部音频的比例；对于匹配任务，准确率为结果字典中值（包括-1）与 ground truth 相同的音频占全部音频的比例。

2. 具体评分标准

本次大作业满分 100 分，占期末总评的 15%，一般情况下，组内成员得分相同。满分 100 分由报告和结果两部分组成，其中报告占 50 分，结果占 50 分。根据三中的任务

描述，结果分由三部分组成：音频分类（15 分）、完全匹配（25 分）、不完全匹配（10 分）。此外，在期末总评满分 15 分的基础上，结果特别优秀的小组，期末总评加 1 分（不超过 100 分）。