音视频场景识别

温海林

June 2023

1 实验介绍

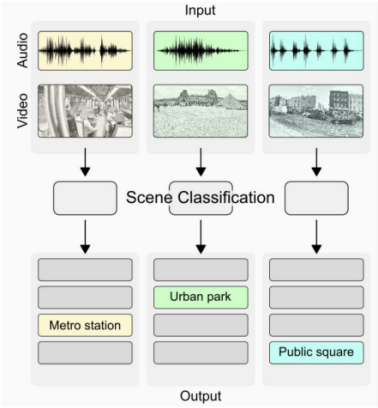

图 1: 模型框架图

Pretrain 阶段通过 openl3 提取 audio 和 visual 的特征向量，训练阶段通过网络得到 video_embedding 和 audio_embedding，将两个向量拼接以后通过全连接层可以得到概率分布图，从而判断属于哪一个类别。

2 early 特征融合和 late 决策融合

early 特征融合是先将 video_embedding 和 audio_embedding 拼接在一起后，输入到一个同一个模型进行分类。late 决策融合是将 video_embedding、audio_embedding 分别输入一个模型进行分类，最后将将各自决策融合。

本实验中只需要将 video_embedding、audio_embedding 分别输入各自的 Output_layer 后，把所得向量相加就行。

3 实验结果

model	val loss	val accuracy
baseline	0.495	0.8078
late 决策融合	0.483	0.8106
late 决策融合 +rate 调整	0.435	0.8336
final model	0.425	0.8417

改为 late 决策融合后，val accuracy 提高到 0.8106，但是只是简单地将决策结果进行相加过于粗暴，video 和 audio 决策结果的重要性可能不相同，因此通过 output = video_output * rate + audio_output * (1 - rate) 计算得到最后的决策结果。调整不同的 rate 值进行比较以后，选择 rate = 0.67，此时 val accuracy 提升到 0.8336 左右。

最后，观察到 train loss 和 val loss 相差较大，训练后期模型过拟合严重，于是调小 lr, 缩小 hidden size, 调整参数之后，最后的模型 val accuracy 位 0.8417。

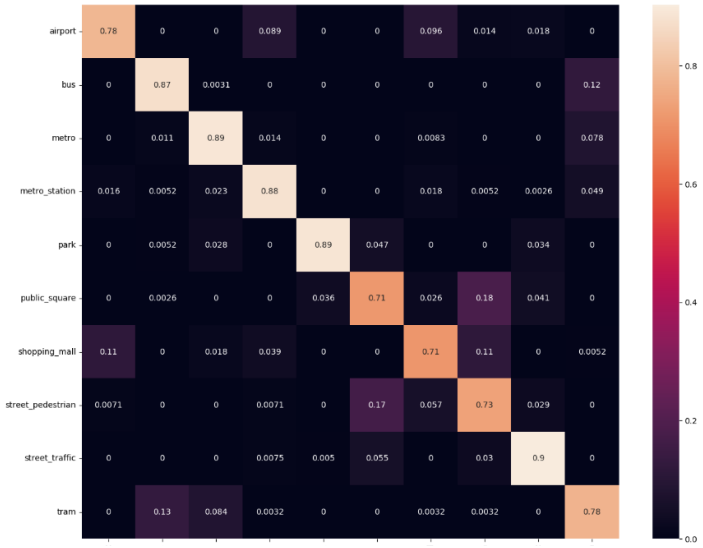

图 2: 实验结果

4 实验结果分析

模型在 bus,metro,metro_station,park,street_traffic 上正确率较高，在其他类别上判断成功率较低。观察概率分布图可以发现，实验结果几乎呈现一个对称矩阵，即如果 public_square 误判为 street_pedestrian 率高，那么 street_pedestrian 误判为 public_square 的概率也较高，说明两者 audio 和 video 数据存在较高相似性。而那些判断成功率较高的类别和其他类别场景、音频相似性较低。因此当一个类别的声音或场景自身特点较为突出的时候，不容易造成场景误判。