音视频场景识别项目报告

520030910342 柳纪宇

1. 概述

音视频场景识别是一种多模态表征学习任务，该任务以音频与视频信息作为输入，将预测得到的唯一场景作为输出。在该任务中，我了解了多模态任务中表征获取与模态融合的各种方式，尝试了早期融合、晚期融合、混合融合和注意力模型四种融合机制，并针对它们的实验结果进行了分析。

2. 表征获取与模态融合

多模态学习的一大难题在于如何表征各模态中的信息和如何将多模态的特征联合起来：更好的表征方式使得网络的输入能够包含更多的信息，也意味着这些信息能够被网络更好地学习；而多模态融合能够使得各模态间共享的信息得到增强，并使各模态的信息达到互补的效果。接下来我将分别介绍多模态学习中不同表征的获取方式和融合方式。

2.1. 表征获取

表征即将原始数据转化成网络输入的过程，好的表征方式能够使输入模型的数据平滑、稀疏、易于学习。多模态学习的原始数据一般有文字、音频、视频等类型，这些表征之间的组合和协同处理也是多模态模型成功的关键。多模态表征主要分为两种：联合表征 (joint representations) 和协同表征 (coordinated representations)。

联合表征 联合特征（见图1[1]）表示的主要理念在于将不同模态的原始信息映射到同一个特征空间。这种表示方法通常通过对多模态数据进行特征提取和融合，形成一个统一的表示。这种表征方式的优点是可以捕捉到不同模态之间的关联和互补信息，从而提高学习任务的性能。

协同表征 协同表征（见图2[1]）的主要理念是独立地将每个模态的原始信息映射到各自的特征空间，但保证映射后的向量满足一定的相关性约束（如线性相关）。协同表征的主要优点在于它允许不同模态信息的不同表征，无需寻找一种适合所有模态的共同特征空间。

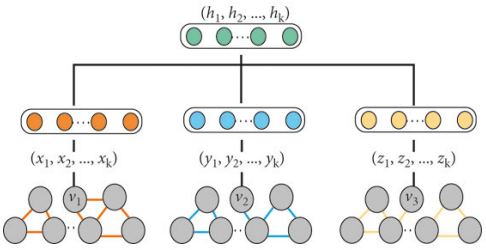

图 1: 联合表征

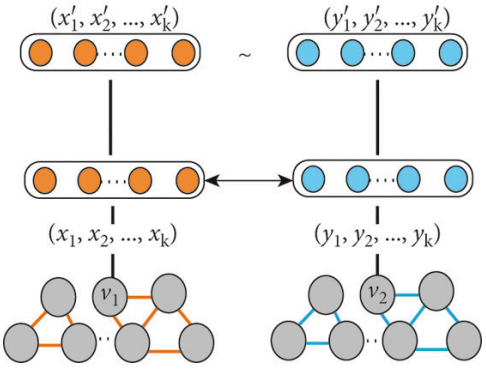

图 2: 协同表征

2.2. 特征融合

如何结合多个模态的特征也是多模态学习领域的一大挑战，好的特征融合方式能够实现缺失信息的补全，增强模型的预测能力和抗干扰能力。目前主流的融合方式有早期融合 (early fusion)、晚期融合 (late fusion)、混合融合 (hybird fusion) 三种。

早期融合 早期融合是指在特征提取阶段就将不同模态的数据融合在一起，形成一个统一的表示。这种方法通常通过将不同模态的原始数据或低级特征进行拼接或加权求和来实现。早期融

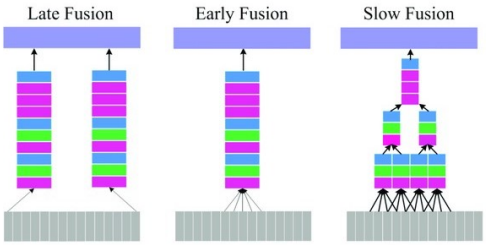

图 3: 各多模态特征融合方式效果图

合能够捕捉到跨模态的关联信息，但可能在融合过程中丢失某些模态特有的信息。早期融合适用于模态间关联较强且相互补充的场景。

晚期融合　晚期融合是指在模型训练的后期阶段将不同模态的特征进行融合。通常，这种方法首先独立地为每个模态训练一个模型，然后在输出层或决策层将各模态的结果进行整合。晚期融合可以保留各模态的独立信息，但可能较难捕捉到跨模态的关联信息。晚期融合适用于模态间关联较弱或各模态特征具有独立价值的场景。

混合融合　混合融合是指模型在各个阶段对多模态特征进行逐级融合的一种方式。通过在不同层级上依次进行特征融合，模型既能够捕捉多模态之间的关联信息，又能不丢失各模态间的独立信息。

以上三种融合形式的可视化效果图见图3[2]，该图从左至右分别为晚期融合、早期融合、混合融合。

3. 实验流程

在本次实验中，我尝试了仅使用音频特征 (audio only)、仅使用视频特征 (video only)、早期融合 (early fusion)、晚期融合 (late fusion)、注意力模型 (attention model) 一共五种模型结构进行音视频场景识别。

3.1. Audio Only

Audio Only 模型由一个音频特征嵌入层和一个 MLP 输出层组成（如图4所示）。

使用该模型得到的训练 loss 变化情况见图5，在 epoch 6 时得到最佳模型，该模型下的 cv loss 为 **0.81**、cv accuracy 为 **0.69**。使用该模型在验证集上进行音视频场景预测，得到的平均准确率为 **0.701**，验证误差为 **0.788**，预测结果见图6。

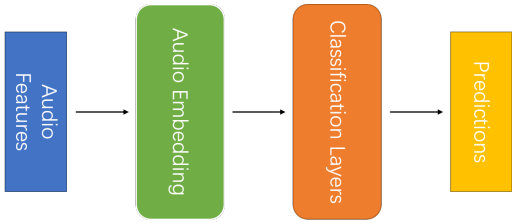

图 4: *Audio Only* 模型结构图

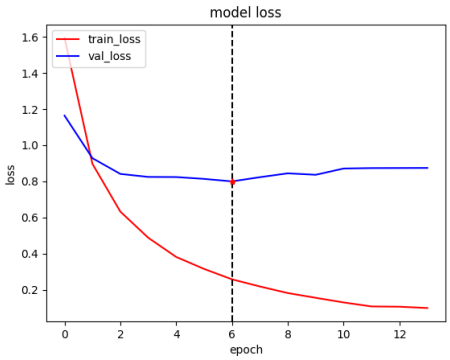

图 5: *Audio Only* 模型训练过程 *loss* 变化图

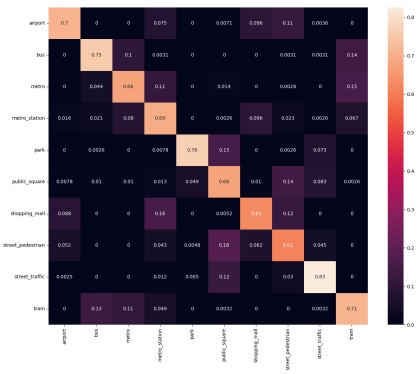

图 6: *Audio Only* 模型预测结果图

3.2. Video Only

Video Only 模型由一个视频特征嵌入层和一个 MLP 输出层组成（如图7所示）。

使用该模型得到的训练 loss 变化情况见图8，在 epoch 3 时得到最佳模型，该模型下的 cv loss

为 **0.78**、cv accuracy 为 **0.66**。使用该模型在验证集上进行音视频场景预测，得到的平均准确率为 **0.671**，验证误差为 **0.894**，预测结果见图9。

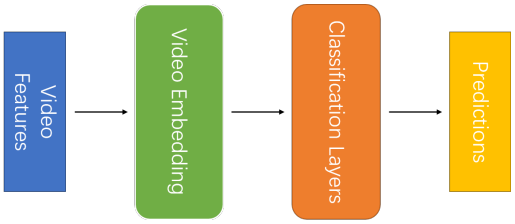

图 7: *Video Only* 模型结构图

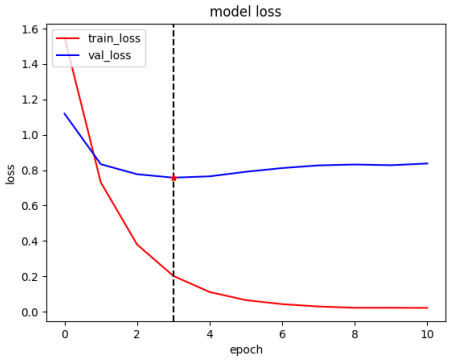

图 8: *Video Only* 模型训练过程 *loss* 变化图

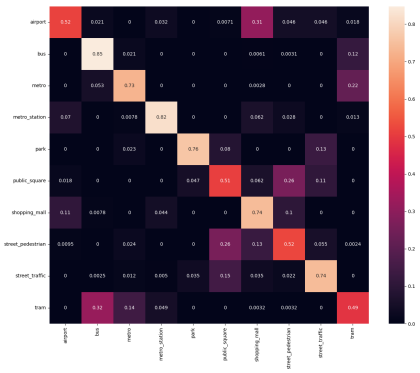

图 9: *Video Only* 模型预测结果图

3.3. Early Fusion

Early Fusion 模型接收音频特征和视频特征作为输入，并为它们分别设计了音频嵌入层和视频嵌入层，在经过嵌入层后将音频和视频的深层特征进行合并，最终输入 MLP 输出层得到预测结果（如图10所示）。

使用该模型得到的训练 loss 变化情况见图11，在 epoch 4 时得到最佳模型，该模型下的 cv loss 为 **0.47**、cv accuracy 为 **0.80**。使用该模型在验证集上进行音视频场景预测，得到的平均准确率为 **0.815**，验证误差为 **0.503**，预测结果见图12。

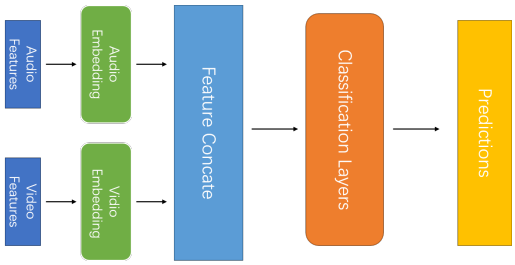

图 10: *Early Fusion* 模型结构图

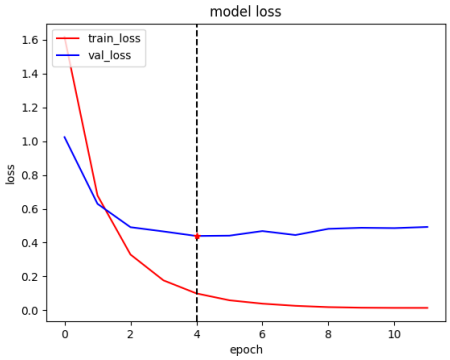

图 11: *Early Fusion* 模型训练过程 *loss* 变化图

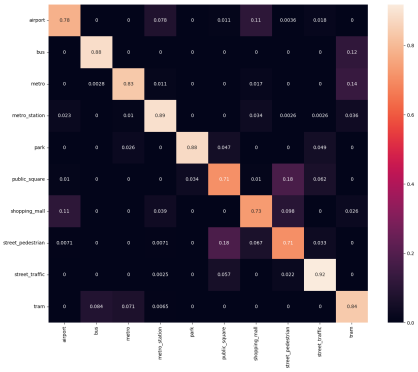

图 12: *Early Fusion* 模型预测结果图

3.4. Late Fusion

Late Fusion 模型同样接收音频特征和视频特征作为输入并分别设计了音频嵌入层和视频嵌入层。与 Early Fusion 不同的是嵌入得到的音频深层特征与视频深层特征都将通过各自的分类器得到预测结果，接着再将二者进行加权求和得到最终预测结果（如图13所示）。

使用该模型得到的训练 loss 变化情况见图14，在 epoch 5 时得到最佳模型，该模型下的 cv loss 为 **0.45**、cv accuracy 为 **0.81**。使用该模型在验证集上进行音视频场景预测，得到的平均准确率为 **0.798**，验证误差为 **0.529**，预测结果见图15。

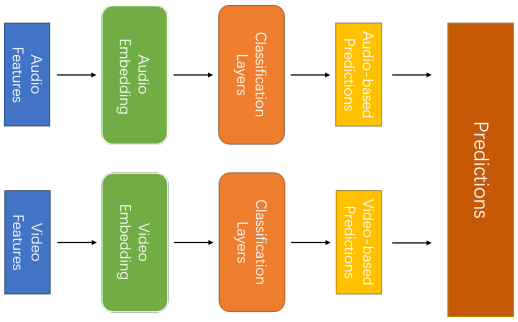

图 13: *Early Fusion* 模型结构图

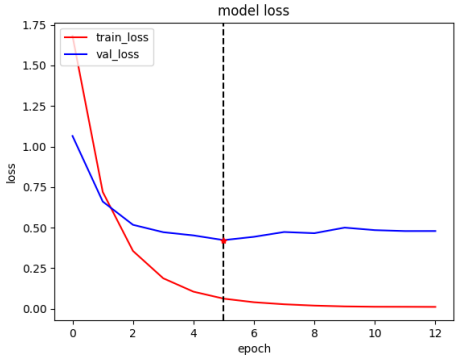

图 14: *Late Fusion* 模型训练过程 *loss* 变化图

3.5. Hybrid Fusion

Hybrid Fusion 模型保留了两条不同的特征融合道路：早期嵌入特征层面的融合和晚期预测结果层面的融合。具体来说，音频特征和视频特征在经过嵌入层之后先进行一次早期融合，接着将该融合张量、音频嵌入特征和视频嵌入特征分别输入分类器中得到预测结果，最后再对三者加权求和得到最终预测结果（如图16所示）。

使用该模型得到的训练 loss 变化情况见图17，在 epoch 4 时得到最佳模型，该模型下的 cv loss 为 **0.50**、cv accuracy 为 **0.77**。使用该模型在验证集上进行音视频场景预测，得到的平均准确率为 **0.798**，验证误差为 **0.529**，预测结果见图18。

3.6. Attention Model

在该模型中，我为音频特征与视频特征的融合引入了注意力机制。除此之外，我还将原先的全连接网络结构改成了基于 LSTM 循环神经网络的编码器-解码器结构。具体实现方式为：先将音频特征和视频特征分别输入 LSTM 编码器得到编码器输出和隐层信息，接着将拼接后的编码器输出和隐层输入注意力层得到上下文向量，然后将该向量与隐层信息输入 LSTM 解码器中得到解码器输出，最后将其输入线性分类器中得到最终预测结果。

使用该模型得到的训练 loss 变化情况见图19，在 epoch 1 时得到最佳模型，该模型下的 cv loss 为 **0.76**、cv accuracy 为 **0.69**。使用该模型在验证集上进行音视频场景预测，得到的平均准确率为 **0.759**，验证误差为 **0.637**，预测结果见图20。

3.7. Attention Model2

由于上述注意力模型仅考虑了融合后的特征的注意力分数，而没有考虑音频特征和视频特征各自内部的注意力关系，所以我又设计了一种结合了三种注意力关系的模型（称之为 MultiModelAttention2）。该模型的基本结构由视频编码器、音频编码器、音频注意力层、视频注意力层、多模态注意力层、ChildSum 层、解码器和输出层组成。具体实现方式为：先将音频特征和视频特征分别输入各自的编码器得到编码器输出，接着将音视频编码器输出和隐层状态输入 ChildSum 层得到解码器隐层状态与细胞状态。然后将音视频各自

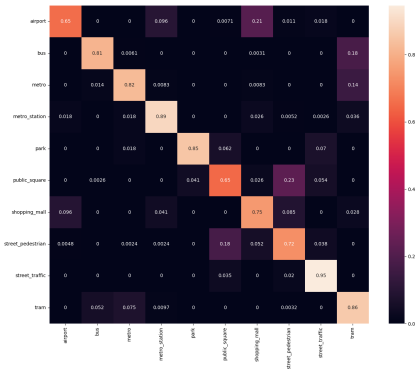

图 15: *Late Fusion* 模型预测结果图

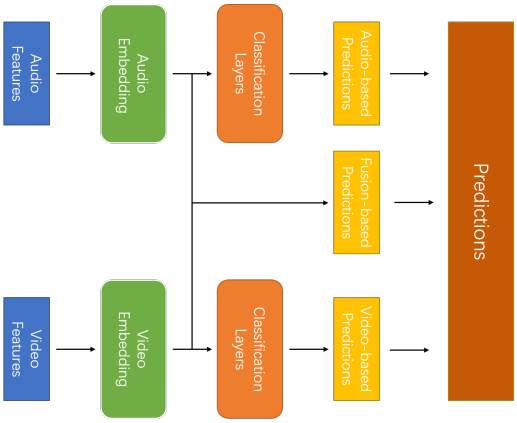

图 16: *Hybrid Fusion* 模型结构图

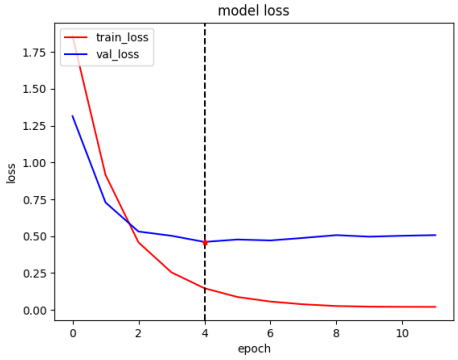

图 17: *Hybrid Fusion* 模型训练过程 *loss* 变化图

的编码器输出和解码器隐层状态输出音视频各自的注意力层中得到音视频上下文向量，对其进行拼接得到多模态上下文。将该上下文输入解码器，再将解码器输出输入输出层得到最终的预测结果。

使用该模型得到的训练 loss 变化情况见图21，在 epoch 1 时得到最佳模型，该模型下的 cv loss 为 **0.96**、cv accuracy 为 **0.59**。使用该模型在验证集上进行音视频场景预测，得到的平均准确率为 **0.725**，验证误差为 **0.734**，预测结果见图22。

4. 实验结果分析

根据各模型的预测结果（结果汇总可见表1，其中最优结果使用加粗字体标出），我们可以得出如下结论。

- 使用多模态信息的预测结果明显好于单模态信息。特征融合能够实现多模态信息之间的互补，从而让模型更好地把握到该场景的核心特征。
- 在单模态模型中，Audio Only 模型的预测效果略优于 Video Only 模型，说明在该任务条件下，音频特征对场景预测的作用权重略高于视频特征。
- 在多模态模型中，Early Fusion 模型的预测结果略优于 Late Fusion 模型，说明在模型浅层进行特征合并在该任务下能够达到更好的效果。
- 分析各场景单独的预测结果，我发现 Audio Only 模型对街道交通 (street traffic) 场景的预测结果较好，而对购物商场 (shopping mail)、街上行人 (street pedestrian) 等场景的预测结果较差；而 Video Only 模型在公交车 (bus)、地铁 (metro) 等场景表现较好，而在机场 (airport)、公共广场 (public square)、街上行人和有轨电车 (tram)。这是因为购物商场、街上行人这些场景的音频特征较容易与其他嘈杂场景（如公共广场）混淆，而交通场景的音频特征辨识度较高；同时人流量较

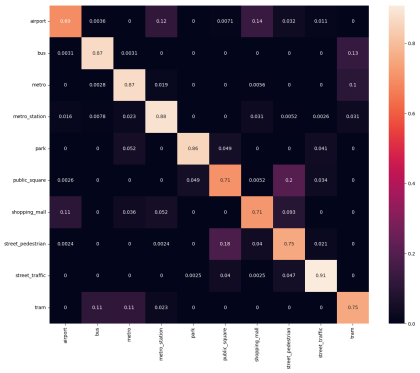

图 18: *Hybrid Fusion* 模型预测结果图

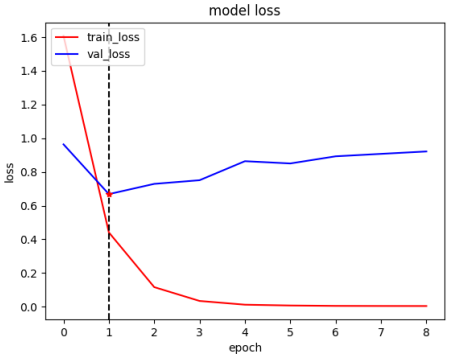

图 21: *Attention2* 模型训练过程 *loss* 变化图

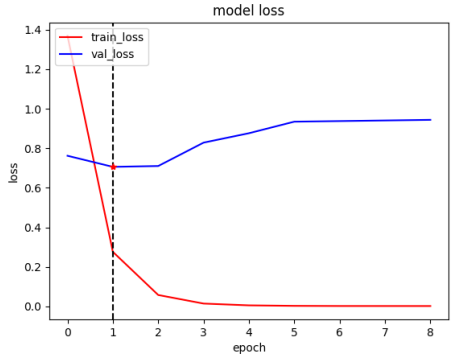

图 19: *Attention* 模型训练过程 *loss* 变化图

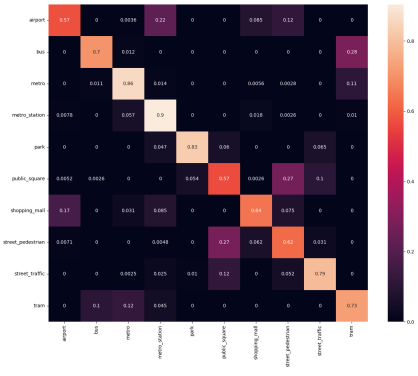

图 22: *Attention2* 模型预测结果图

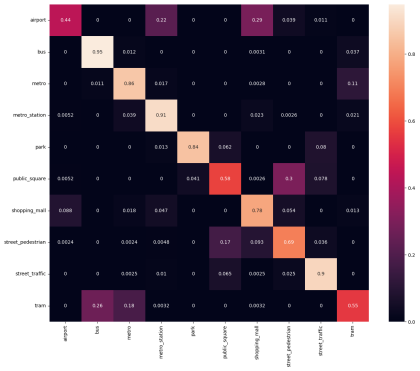

图 20: *Attention* 模型预测结果图

大的公共场所（如机场、街道、广场等）容易相互混淆造成预测困难，而公交车、地铁这些交通工具的视觉特征比较容易被识别。有轨电车预测准确率较低的原因可能是：有轨电车车厢结构与公交车和地铁类似，而在训练集中关于公交车和地铁的视频数量共 3477 个，但有轨电车的视频数量仅 1071 个，故模型有较大的概率将有轨电车场景误预测成公交车或地铁场景。

- 比较 Early Fusion 模型结果与单模态结果，我发现多模态融合在所有场景的预测准确率上都有所提升，其中公交车、地铁站 (metro station)、公园 (park) 和街道交通这四个场景的多模态融合效果最好。其中视频信息在公交车场景中已经相当有效，但音频信息的加入可以提供更丰富的环境背景，比如公交车引擎的声音或者广播的声音等，这可能有助于提高分类准确度。而另外三个场景均属于公共场所，这类场景之间的音视频特征可能非常复杂且相似，如果仅有音频特征则无法很好地区分不同场景的嘈杂声，如果仅有视频特征则难以将它们与较为类似的室内或室外场景区别开。在这种场景上多模态融合就很好地对单模态的不足之处进行了弥补，通过对跨模态特征的捕捉和各模态之间的互补，多模态模型能够更好地处理场景中复杂的音视频信息，从而达到更好的预测效果。

- Hybrid Fusion 模型预测准确率的中位数相较 Early Fusion 更高，但平均准确率没有 Early Fusion 高。推测可能是因为 Hybrid Fusion 模型在通过在早期和晚期都引入特征融合提高了一些准确率较低的场景的准确率；但对于一些本来准确率就比较高的场景，Hybrid Fusion 反而使得 Early Fusion 中获取地较好的一些特征得到了稀释。

- Attention 模型的预测结果有着两极分化的现象，一方面我注意到在公交车、地铁站和道路交通场景中，Attention 模型有着所有模型中最好的效果；而另一方面，在机场、公共广场和有轨电车场景中，Attention 模型的预测效果又是所有模型中最差的。我推测这可能是因为我在实现 Attention 模型时使用的是先得到音视频编码器输出再将其拼接在一起求解注意力的方式，所以如果音视频特征的原始输入不平衡，那么注意力分数很可能会在两个模态上有较大的差距，从而导致性能的下降。

模型结构	Val Accuracy	Val Loss
Audio Only	0.701	0.788
Video Only	0.671	0.894
Early Fusion	**0.815**	**0.503**
Late Fusion	0.798	0.529
Hybrid Fusion	0.803	0.537
Attention	0.759	0.637

表 1: 各模型验证集预测结果汇总表

- 使用了更复杂的注意力机制的 Attention2 模型同样有着两级分化现象，且其总体准确率比 Attention 模型还要低。因此我推测，使用注意力机制的模型在预测结果上表现不佳除了原始输入不平衡的问题外，还可能是模型整体结构过于复杂，增大了模型过拟合的可能性。

5. 总结

在本次实验中，我了解了多模态任务的各种表征获取和模态融合方法，使用了 Early Fusion, Late Fusion, Hybrid Fusion, Attention 四种不同的模型进行了音视频场景识别任务的训练。其中，Early Fusion 模型在该数据集上的表现最好。

注：本报告仅包括模型实现和结果分析，项目结构与运行的相关内容见 readme.md。

6. 参考文献

[1] Nianwen Ning et al. “Nonlinear Structural Fusion for Multiplex Network”. In: *Complex.* 2020 (Jan. 2020). ISSN: 1076-2787. DOI: 10.1155/2020/7041564. URL: https://doi.org/10.1155/2020/7041564.

[2] Chongyang Wang et al. “Micro-attention for micro-expression recognition”. In: *Neurocomputing* 410 (2020), pp. 354–362. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2020.06.005. URL: https://www.sciencedirect.com/science/article/pii/S0925231220309711.