数据挖掘作业 5

1. 任务内容

本任务中，我主要调研了多模态融合工作不同表征获取的方式、模态融合的方式。分析对比了多模态和单模态性能的区别，理解了蛋模态和多模态的应用。并尝试了不同的特征融合方式，尝试替换特征、修改模型、优化参数，以改进模型性能。

2. 多模态特征提取方式

多模态任务中会用到多种形态的数据，针对不同模态的数据会有不同的特征提取方式。可以使用人工设计的特征来进行特征提取，也可以使用预训练模型作为特征提取的手段。

2.1. 文本特征提取

文本特征的提取主要有以下的方法：

词袋模型（Bag-of-Words）：将文本表示为词的集合，忽略词序和语法结构，仅考虑词的出现频率或存在与否。

TF-IDF（Term Frequency-Inverse Document Frequency）：结合词频和逆文档频率，用于衡量词在文本中的重要性。

词嵌入（Word Embedding）：使用预训练的词向量模型（如 Word2Vec、GloVe、BERT 等）将词映射到低维的连续向量空间，捕捉词之间的语义关系。

文本向量化（Text Vectorization）：将文本转换为数值向量表示，可以使用词袋模型、TF-IDF 等方法，也可以使用深度学习模型（如循环神经网络、卷积神经网络）进行文本编码。

2.2. 图像特征提取

图像特征的提取主要有以下的方法：

传统的特征提取方法：使用计算机视觉领域的特征提取方法，如 SIFT、HOG、LBP 等，从图像中提取局部或全局的特征描述符。

深度学习模型提取图像特征：使用预训练的 CNN 模型（如 VGGNet、ResNet、Inception 等）提取图像的高级特征表示。

2.3. 音频特征提取

音频特征提取主要有以下的方法：

短时傅里叶变换（STFT）：将音频信号分解为时频域表示，获取音频在不同时间和频率上的能量分布。

梅尔频率倒谱系数（MFCC）：将音频信号转换为梅尔刻度的频谱表示，捕捉音频的频率特征，并使用倒谱系数表示音频的特征。

声谱图（Spectrogram）：将音频信号转换为时频图像，用于表示音频在不同时间和频率上的能量分布。

以上是比较传统的声学特征提取方法。也可以使用深度学习模型、使用卷积神经网络（CNN）、循环神经网络（RNN）或 Transformer 等深度学习模型，将音频信号转换为低维的稠密向量表示，以此来提取特征。

2.4. 本次任务中的特征提取

本次任务中我们使用的是 OpenL3 进行特征的提取。OpenL3 使用神经网络对多模态数据进行特征的提取。

音频特征的提取是先将音频信号转化为频谱图，然后通过卷积网络来提取固定步长的特征向量。

图像特征也使用了更深的卷积网络来提取，并且使用了大数据集进行训练。

3. 多模态特征融合方式

多模态的特征融合方式主要有早期融合、晚期融合、注意力机制融合等。

3.1. 早期融合

早期融合是指在输入模态之间进行融合，将不同模态的特征或表示直接合并成一个单一的特征向量或表示。在早期融合中，模态之间的信息交互发生在模型的输入层之前。例如，将图像和文本的特征向量连接在一起，形成一个更大的特征向量，然后将其输入到模型中进行训练和推理。

早期融合的优点是可以利用不同模态的特征进行联合训练，从而更好地捕捉模态之间的相关性。可以减少模型的计算复杂度，因为融合后的特征维度较低。

早期融合的缺点是可能会导致信息冗余，因为不同模态的特征可能包含大量相似的信息。对于不同模态之间的异构性较大的情况，早期融合可能无法充分利用每个模态的独特信息。

3.2. 晚期融合

晚期融合是指在模型的较高层次或输出层之前进行融合，将不同模态的特征或表示分别输入到各自的模型中进行处理，然后将它们的输出进行融合。在晚期融合中，模态之间的信息交互发生在模型的中间或最后的层次。例如，将图像和文本分别输入到各自的卷积神经网络和循环神经网络中，然后将它们的输出特征向量合并，并通过全连接层进行最终的预测。

晚期融合的优点是可以充分利用每个模态的独特信息，因为每个模态都经过独立的处理和学习。可以适应不同模态之间的异构性，因为每个模态可以使用适合自身的模型结构。

晚期融合的缺点是可能会导致模型的计算复杂度增加，因为需要独立处理每个模态。可能会忽略模态之间的相关性，因为融合发生在较高层次或输出层之前。

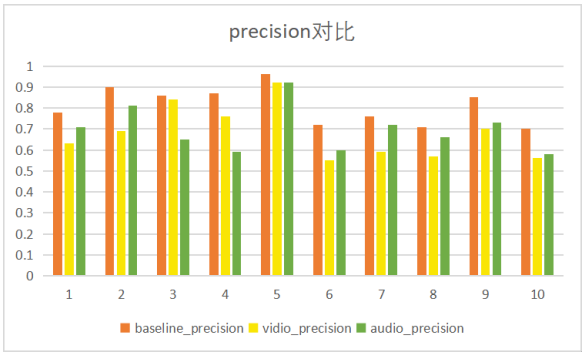
3.3. 注意力机制融合

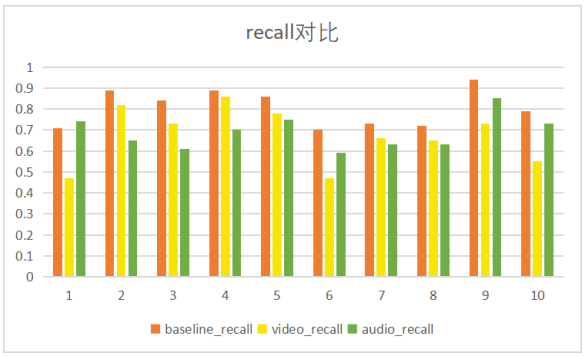
注意力机制融合是一种基于注意力机制的融合策略，它允许模型自动学习不同模态之间的关联程度，并根据这些关联程度对不同模态的特征进行加权融合。在注意力机制融合中，模型可以学习到每个模态对于特定任务的重要性，并根据这些重要性对模态特征进行加权融合。例如，使用注意力机制来计算图像和文本之间的注意力权重，然后将这些权重应用于对应的特征向量，以实现模态之间的自适应融合。

注意力机制融合的优点是可以自适应地学习模态之间的关联程度，从而更好地融合不同模态的特征。可以根据任务需求，动态地调整不同模态的重要性。

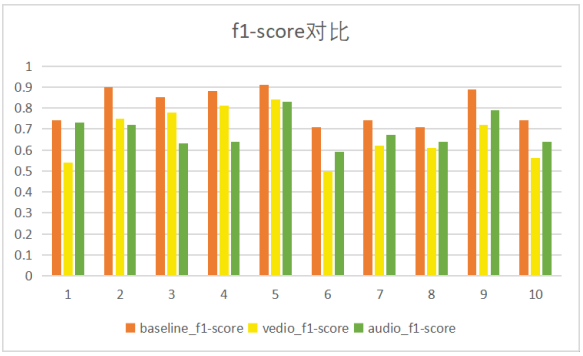
缺点是可能会增加模型的计算复杂度，因为需要计算注意力权重。对于数据量较小或模态之间关联性较弱的情况，注意力机制可能无法有效学习到合适的权重。

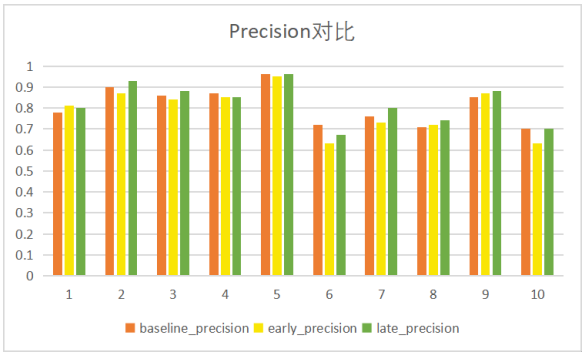
4. 单模态和多模态结果对比

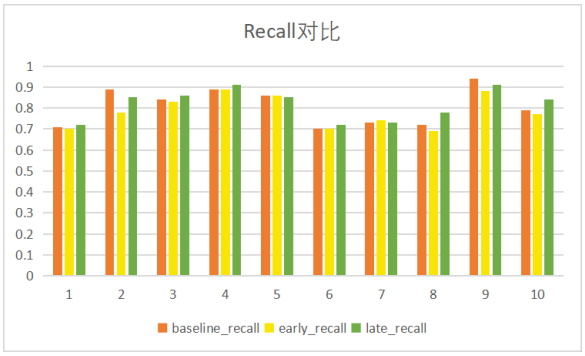




从左到右依次是 airport、bus、metro、







metro_station、park、public_square、shopping_mall、street_pedestrian、street_traffic、tram

我发现使用多模态融合后，效果会比单模态都有所提升。

仔细观察后，我发现模型在 airport 类中的效果没有很明显，甚至 recall 的指标，audio 会比多模态更好。我认为可能是以下情况导致的：

1、数据不平衡：多模态数据集中 airport 场景的样本数量较少，而其他场景的样本数量较多，模型可能更倾向于学习其他场景的特征，而在 airport 场景中的多模态融合效果不明显。

2、特征选择不当：如果在 airport 场景中选择的特征与其他场景的特征不够匹配或不具有区分性，可能导致多模态融合效果不佳。

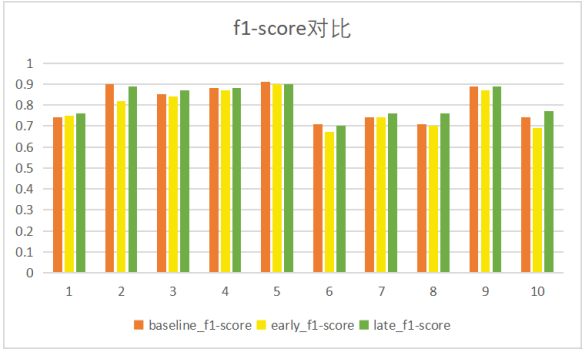
3、数据本身存在问题：多模态数据集中的图像和文本可能存在噪声或错误，这可能导致多模态融合的效果不佳。确保数据集中的图像和文本与音频数据质量相当重要。

我因此尝试更换了更深的模型，效果有所提升但不明显。

我认为也可以尝试更换注意力机制融合的方式，我实践了一部分，最终没能正确运行。

5. 决策融合方式

我实现了 early 特征融合和 late 决策融合两种方法，略微调整了一些参数，运行得到的结果和 baseline 的对比如下：



观察发现三种方式差别不大，都能取得较好的效果，细致来看 late 方式更好一些。

原因分析我认为是晚期融合允许每个模态使用适合自身的模型结构进行处理，这样可以更好地发挥每个模态的表达能力。通过独立地处理每个模态，模型可以更好地学习到每个模态的特征表示，从而提高整体性能。