模型架构

图1.本项目提出的方法的总体架构

本项目的目标时通过给定一个包含人们直接对着镜头说话的vlog数据集，训练一个模型来估计视频中的主体是否患有抑郁症。在这种设定下，抑郁症检测相当于视频级别的分类。然而，如果没有不切实际的大量标注数据，直接从像素进行高级心理疾病的分类是不可行的。因此，在我们的工作中，我们仅考虑了高级别的非语言线索，并丢弃了任何文本信息以消除对话主题的偏差[t80]。大多数抑郁症数据集的收集方式是基于直接提及的诊断[t86,88]，通过仅使用非语言线索，模型的性能更能反映其在现实环境中的表现。非语言线索是使用最先进的预训练模型[t54,70,15,81,90,89,改]提取的，随后进行处理以进行分类。图1提供了我们从视频中进行抑郁症分类的流程的高级概述。

正式地来讲，给定一个带有抑有症标签的视频数据集 ，其中是视频，是其对应的标签，目标是找到模型的最优参，以最小化整个数据集上的平均交叉熵损失：



其中，表示交叉熵损失函数。

窗口采样

由于现实中的视频和vlog时长和帧率差异很大（见图3），有些视频甚至超过30分钟，直接完整处理这些视频在计算上是不可行的。因此，我们对每个视频随机采样固定时长（以秒为单位）的时间窗口  ，其中  是视频中的随机时间索引。时间窗口可以覆盖不同数量的总帧数，具体取决于每个视频的帧率。音频以每秒100帧采样［27］，而D－Vlog数据集［87］中的视频帧率在每秒6到30帧之间。通过窗口采样方法，我们假设视频中的心理健康信息在时间上是恒定的。

模态提取与编码

对于每个视频，我们考虑了多个语义模态  ，这些模态是通过冻结的预训练模型从视频的音频和视觉信息中逐帖提取的。因此，视频在所有时间串联模态中的总帧数为  。由于每个模态的维度  不同，我们使用一个可学习的模态编码器  对每个模态的输出进行统一化，使其具有相同的维度  。因此，窗口  和模态  的特征向量为  ，其中  。关于每个模态编码器  的详细信息见第3．2节。此外，每个模态提取器可以指示某一帧中是否存在该模态，从而允许我们构建一个二进制存在掩码  。该存在掩码随后用于最终Transformer编码器［73］的注意力掩码中。在实践中，对于现实中的视频，我们首先提取并缓存每个视频的所有模态，然后对窗口进行采样，确保主体在窗口的至少  的帧中出现。这是因为一些视频是业余制作的，噪声很大，主体的身体，手和脸并不总是可见；例如，有些人在开车时说话，方向盘遮挡了摄像头，或者主体走出了画面。

分数位置嵌入

由于采样率不匹配，我们通过使用分数位置嵌入［33］  来对齐基于音频和视频的模态帧。对于视频，我们构建一个大小为  的位置嵌入矩阵，对应于采样率最高的模态（通常是音频）。然后，模态  的位置根据比率  均匀索引。分数位置嵌入基于正弦位置嵌入［73］，并且在训练期间不学习。跨模态的帧在时间上的相同位置具有相同的位置嵌入，与采样率无关（见图2）。更多细节请参考Harzig等人［33］。

特征向量处理

修改后的特征向量为  。最后，我们将拼接后的修改特征向量  与每个对应的存在掩码  一起输入Transformer编码器网络［73］，以获得预测  ，其中  。最终的优化公式可以描述为：其中， L 表示交叉摘损失函数。