[数据集] D-vlog: Multimodal Vlog Dataset for Depression Detection

1. Abstract

   D-Vlog收集YTB的961个vlog，建立**现实场景**下，**非言语**的抑郁症患者**声学和影像交叉**数据集。开发了**基于交叉注意力机制**的**多模态**DL模型，识别检测抑郁症。填补过去数据集过多在实验室环境下的空白

2. Intro

   （抑郁症及其治疗略）非言语的信号（表情、行为、音量）对区分抑郁症患者和普通人相当有用。但目前自动判别存在困难，因为：

   - 主要因为隐私问题，公开的数据集很少
   - 大部分数据集是在实验室环境下的采访，难以捕捉平常的行为特征

   本文作者收集了含相关关键词（“抑郁”“日常”）的vlog，基于严格的规则手工标注得到数据集。此外，还建立了一个多模态DL模型，使用encoder编码声音和影像序列，并用交叉注意力机制产生多模态表示，超越已有的其他模型。

   作者提到的主要贡献：

   - **数据集公开**（去除vlogger个人信息，匿名数据）
   - （就他们所知）**第一个**使用交叉注意力进行多模态抑郁检测的尝试。确认了其有效性
   - **可泛化**，例如对临床访谈情景也有效

3. Related Work

   a. 抑郁分析检测数据集

      公共数据集很少，主要有：

Dataset	模态（AVT-音视文）	个体数	样本数	内容	标注方式
DAIC-WOZ	AVT	189	189	模拟重性抑郁、PTSD的临床访谈	量表自我报告
Pittsburgh	AV	49	130	临床访谈	临床评估
AViD-Corpus	AV	292	340	歌唱、阅读等的视频	量表自我报告

   b. 使用社交媒体数据的抑郁检测

      对社交媒体数据的检测，有助于获取抑郁个体日常生活的常见行为模式

      过往工作侧重于文本（ta博客文本，图片，tag等）缺乏对视频数据的分析

   c. 多模态融合

      以往多模态融合大多和模型无关；深度学习时代之后，融合步骤大多集成在模型内。本文使用transformer可以有效融合。

      模态融合是复杂的内容，需要更深一步研究学习。

4. Dataset

   a. vlog收集

收集2020.1.1~2021.1.31抑郁类和非抑郁类各2000个youtube视频，提供了搜索关键词（抑郁类含depression）。

b. 标注

招募4位大学生并培训，每人分配2类个500个视频，要求他们标注：

- 是否符合vlog格式（单人之间与镜头说话），移除不符合的视频
- 是否包含当前抑郁症状的描述，移除无疑似抑郁表达和已经恢复者

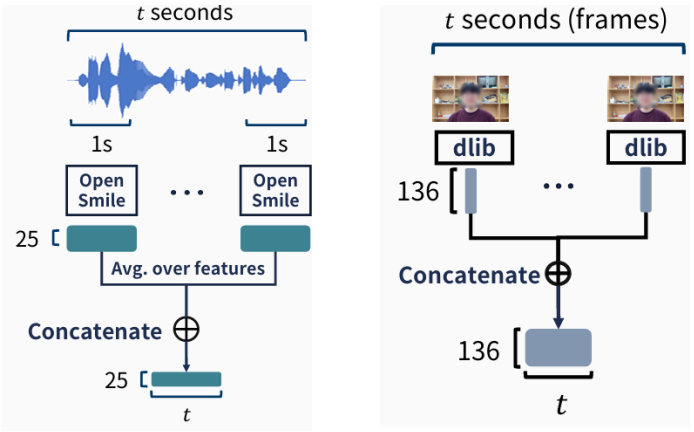
c. 数据统计

	Gender	# Samples	Avg. Duration
Depression	Male	182	583.74s
	Female	373	667.63s
Non-depression	Male	140	438.77s
	Female	266	587.76s

特征：

- 两类vlog都为女性居多，前者与抑郁症患病性别分布符合，后者可能是tag选取的问题
- 视频长度和频道vlog数都有明显长尾效应。视频长度大多小于30min，各频道vlog数大多只有一个降低了因为特定人物而过拟合的可能性

d. 特征提取



(a) Acoustic features (b) Visual features

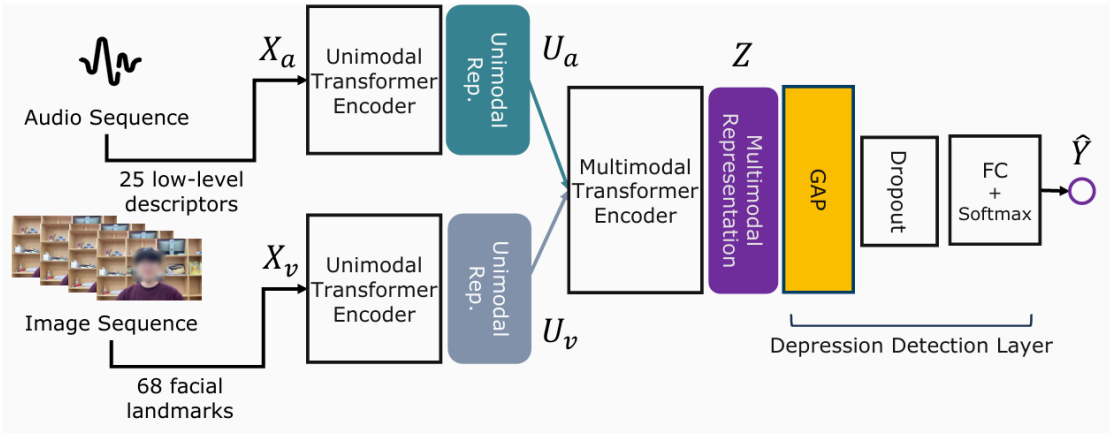
Figure 2: Feature extraction process of (a) acoustic and (b) visual features. We extract 25 acoustic features and 136 visual features for each second.

5. Detection

a. 问题状态

一个二分类问题，给定vlog set $P = \{p_n\}_{n=1}^{|P|}$，每个vlog表示为$p_n = (X_a^n \in \mathbb{R}^{t \times d_a}, X_v^n \in \mathbb{R}^{t \times d_v})$，大X为特征，t为序列长度，d为特征维度，给出是否抑郁的分类

b. 总体架构



c. 单模态Transformer Encoder

允许模型捕捉各个模态的有用特征

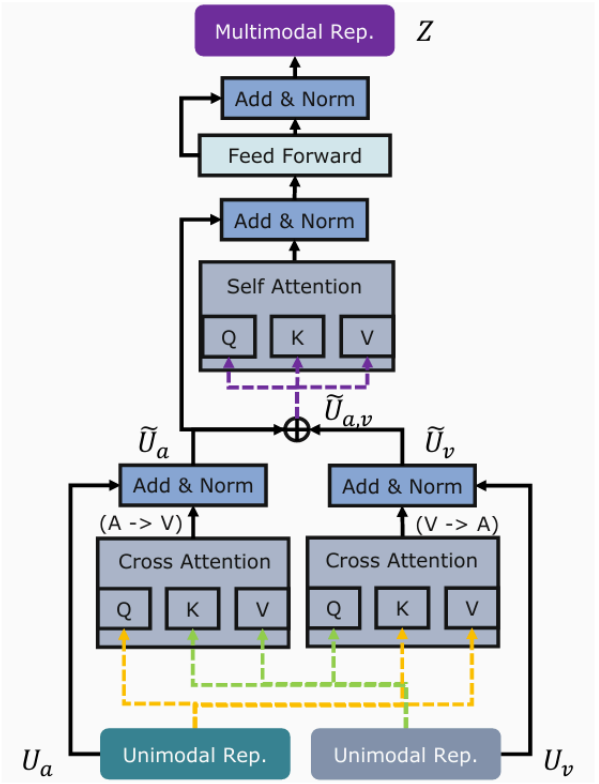
注意：此encoder不限于特定模态，可以直接通过增加encoder融入新的模态

流程：

- 下采样
- 1维卷积层，对局部关系处理
- 位置编码
- 按照经典transformer encoder方式处理

d. 多模态Transformer Encoder

使用交叉注意力机制。有待学习具体内容



e. 抑郁检测层：分别为全局平均池化-Dropout-全连接-SoftMax

$$\hat{Y} = Softmax(\mathcal{F}(Dropout(GAP(Z)))),$$

loss func为交叉熵，前后分别为GT分布和预测概率，0为正常，1为抑郁

$$\text{loss} = \sum_{c \in [0,1]} P(c \mid \mathbf{y}) \log P(c \mid \hat{\mathbf{y}}),$$

6. Exps

具体参数省略

作者进行了多模态融合方式，使用模型的消融试验，以及对不同模态、不同性别数据集的贡献和特征分析，最后，还使用DAIC-WOZ数据集进行交叉语料库试验。

- Fusion Baseline 多模态融合
  使用三种常见方式：矩阵拼接、相加和相乘，代替多模态Encoder
  本文的模型表现最优。注意到三种基准模型中，矩阵相乘表现最好
- Model Baseline 使用模型
  使用诸如逻辑回归LR，支持向量机SVM，随机森林RF等传统机器学习方法，按照其预期输入，进行矩阵展平、拼接等相应操作。具体细节省略。
  实验证明本文模型的性能最高。LR SVM RF等由于不能处理序列信息，劣于设计用于处理序列信息的双向长短期记忆BLSTM和张量融合网络TFN。
- 模态分析
  单独使用音频和影像数据训练，移除多模态encoder部分。
  测试结果表明：
  - 音频携带信息量更大，贡献更大
  - 多模态训练效果优于单模态
- 性别分析
  单独使用男性/女性vlogger的数据训练。
  结果表明：
  - 单男性数据训练效果优于单女性（可能因为男性对抑郁表现的特征更多）
  - 两性数据全部使用效果优于使用单个
  - 无论训练集单性双性，模型对男性预测效果更佳
- 跨语料验证

  使用本文数据集D-vlog(DV)和DAIC-WOZ数据集(DW)，分别作为训练集和测试集

Train	Test	Precision	Recall	F1-Score
DW	DV	60.14	60.38	60.24
DV	DV	65.40	65.57	63.50
DW	DW	62.57	52.63	55.45
DV	DW	69.45	55.26	57.73

结果验证了：

- 本文的DV数据集特征更有用
- 使用生活场景数据的DV训练出的模型，也能泛化到临床访谈场景的数据集DW

7. Conclusion

   本文介绍D-vlog，从包括861人的961个vlog提取非言语的声音和影像信号。同时介绍基于此的，使用交叉注意力机制的多模态transformer模型。它可以根据视频的声音和影像判断人是否患有抑郁，能帮助患病早期的患者及时就医。