NUSTM

南京理工大学 文本挖掘实验室
Text Mining Lab, Nanjing University of Science & Technology

多模态对话中的情感识别评测
技术分享

郑文杰，虞剑飞，夏睿

南京理工大学文本挖掘实验室

2023年7月1日

目录

数据预处理

- 视频预处理
  - 按照标注中给的时间戳，通过OpenCV，切分出无声视频
  - 按照标注中给的时间戳，通过FFmpeg，从原视频中提取音频并切分
  - 利用FFmpeg合并无声视频和音频
- 各模态预处理
  - 文本模态

    将一个对话下的所有utterance通过[SEP]连接起来
  - 语音模态

    将每个utterance的音频送入到Wav2vec2.0模型中，得到768维的词级别语音特征
  - 视觉模态

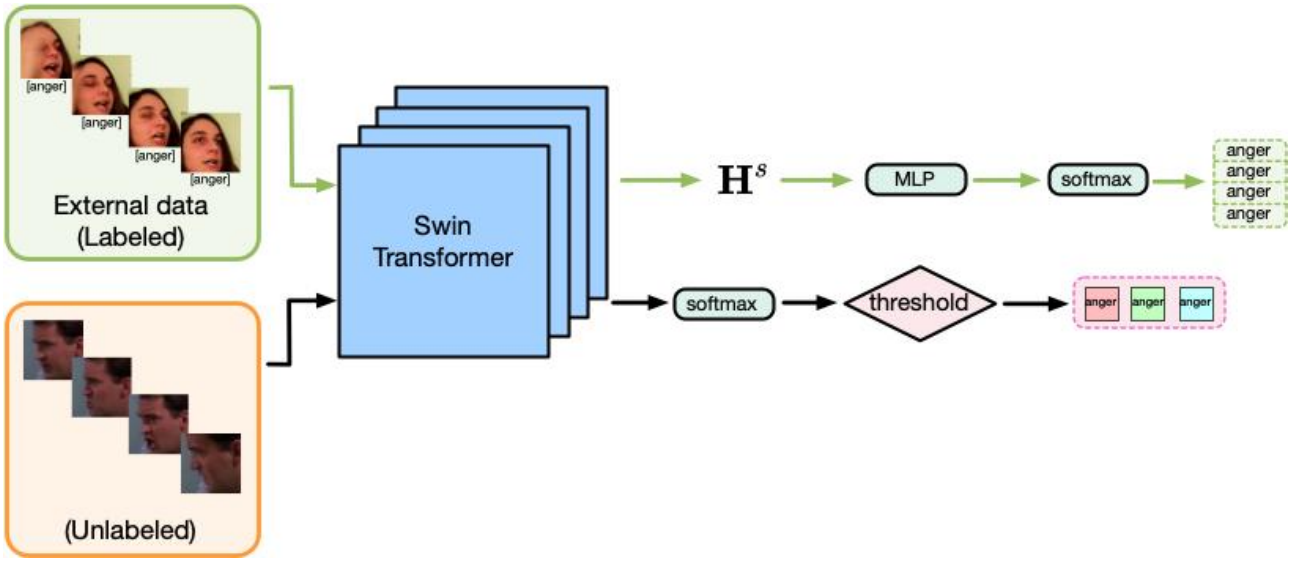
    通过多模态主动说话者检测(Talk-Net)方法，抽取说话人的面部序列

框架提出

- **利用说话人的面部表情辅助多模态情绪识别**
- **两阶段框架**

第一阶段

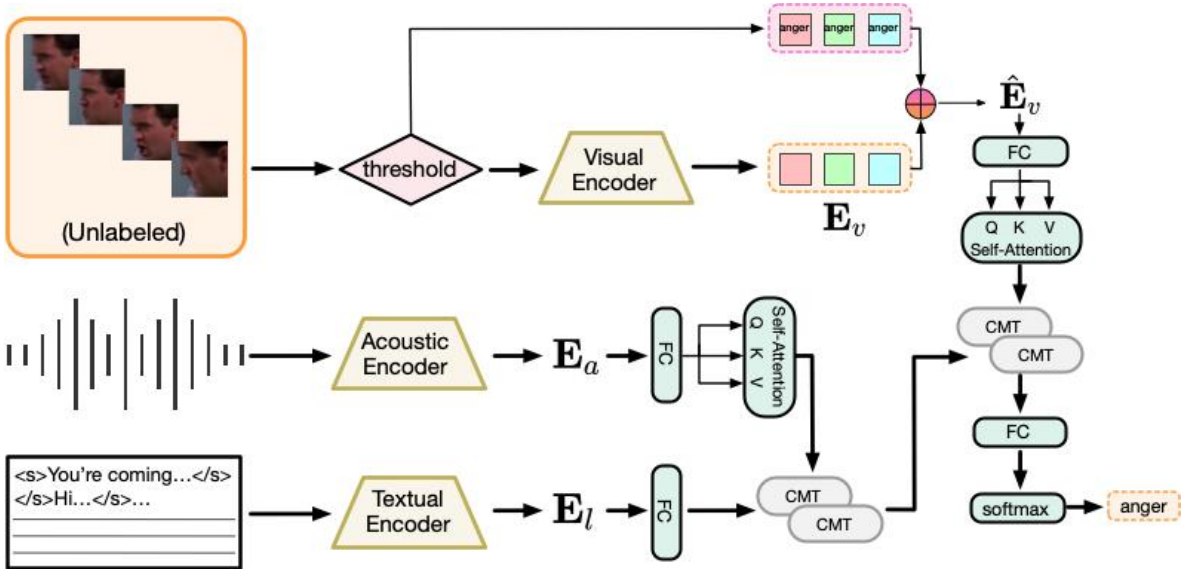
- 利用在Swin-Transformer在做面部表情识别任务的Aff-Wild2数据集上训练一个7分类器
- 将训练好的Swin-Transformer在目标数据集上进行推理，拿到面部表情分布g
- 过滤掉一些表情置信度低的人脸，通过$g \times g^T$



框架提出

第二阶段

- 将过滤后的人脸过InceptionResNetv1模型得到512维特征，再拼接第一阶段拿到的说话人面部表情分布，拿到519维的帧级别视觉特征
- 语音模态和视觉模态分别过Self-Attention Transformer, 拿到各自的utterance级别表示
- 文本模态，通过微调RoBERTa拿到[SEP]位置上的表示作为一个对话下每个utterance的512维文本表示
- 跨模态交互。首先文本和语音过Cross-Modal Transformer (CMT)，得到文本-语音模态交互特征，再与视觉模态做交互，得到最终的多模态特征



实验结果

- 多模态情绪识别任务

提出方案的最佳Macro-F1值为0.4946，取得了本次评测的第四名。

- 本次评测提出方案是我们ACL 2023工作的一个简化版本.

- ACL工作提出的方法在MELD数据集上获得当前的SOTA. 欢迎感兴趣的老师和同学关注我们的工作:

  Wenjie Zheng, Jianfei Yu, Rui Xia and Shijin Wang:
  A Facial Expression-Aware Multimodal Multi-task Learning Framework for Emotion Recognition in Multi-party Conversations. ACL 2023 (Main Conference), long paper.

- 该工作的数据及代码开源，链接为:

  https://github.com/NUSTM/FacialMMT

- 我的邮箱是: wjzheng@njust.edu.cn，欢迎有问题与我交流。

Code