评测任务一：多模态对话中的情感识别挑战赛

基于投票方式的多模态情感识别方法

AI4AI团队：李启飞 王聪 任一鸣 王栋 **高迎明** 李雅

单位：北京邮电大学-人工智能学院

时间：2023年07月01日　　地点：陕西，西安

提纲 OUTLINE

1 研究背景及意义

2 方案设计

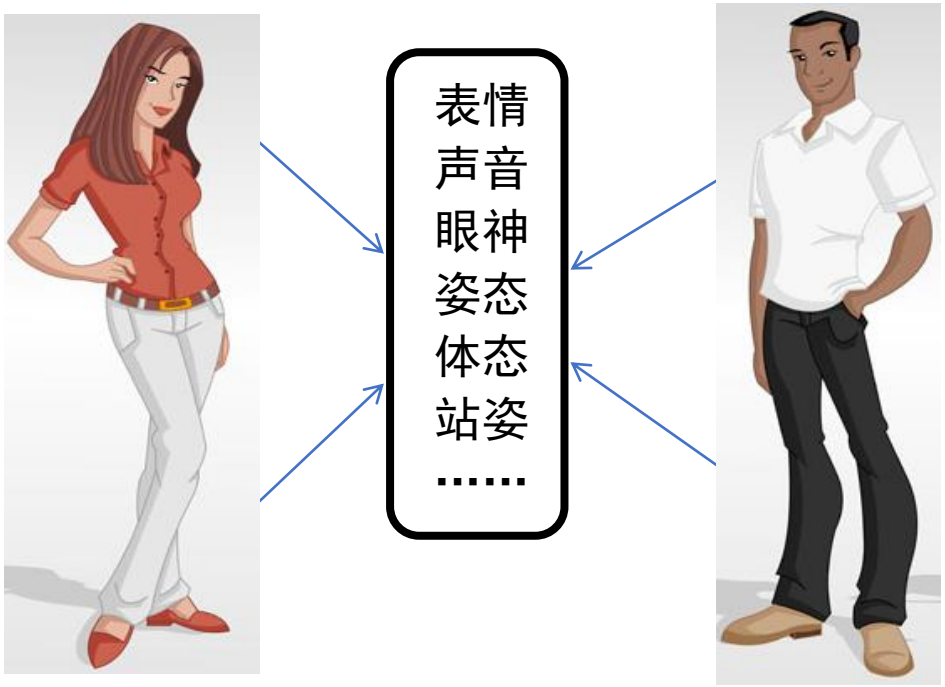
3 方案实现

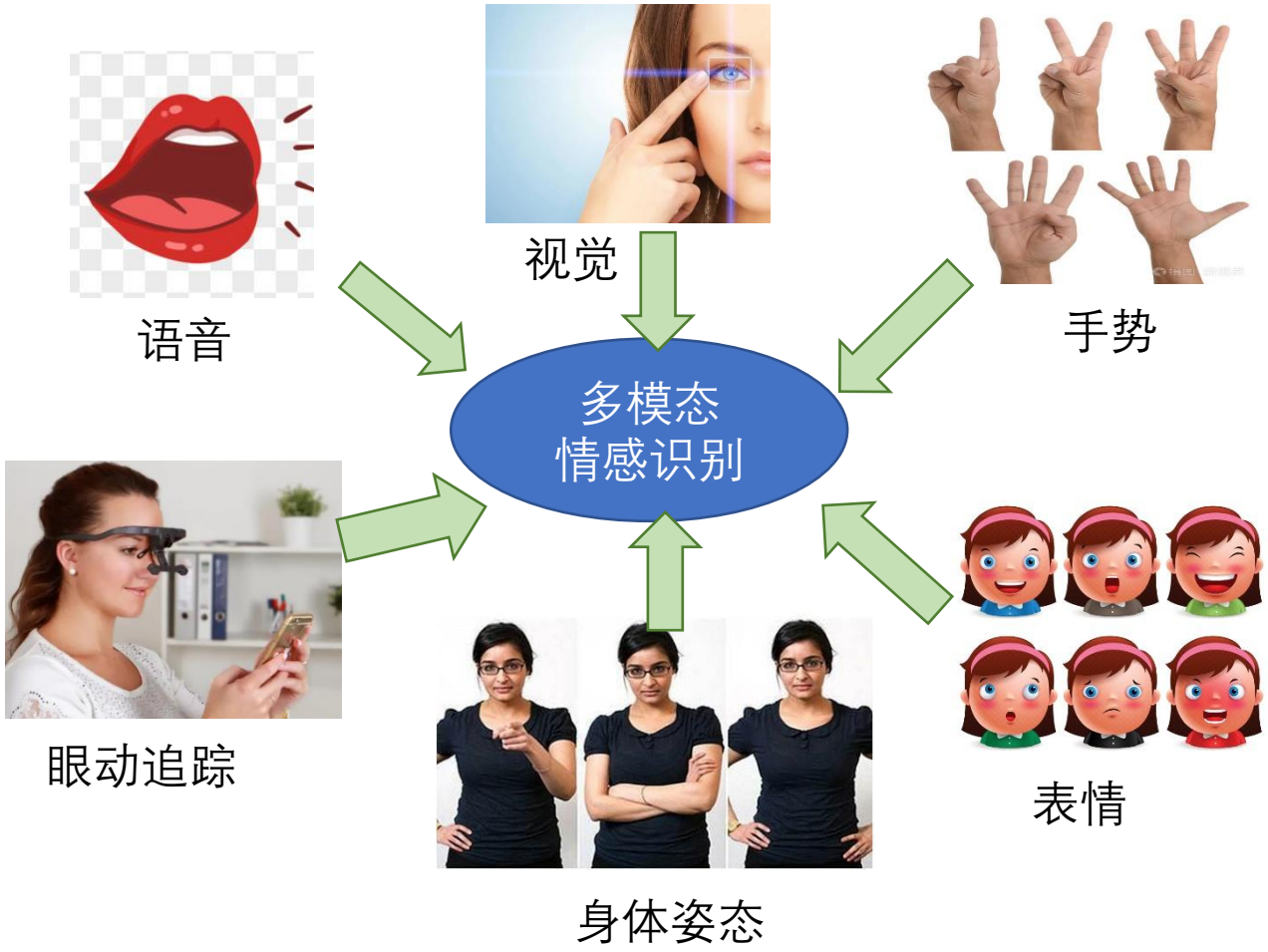
4 数据集及预处理

5 方案评估与总结

研究背景

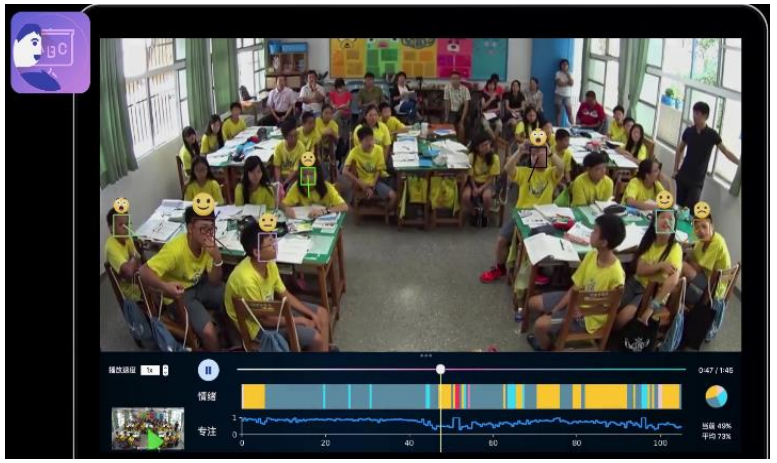
- 多模态情感识别
  - 情感是交流的重要组成部分
  - 多种模态传达丰富的情感信息





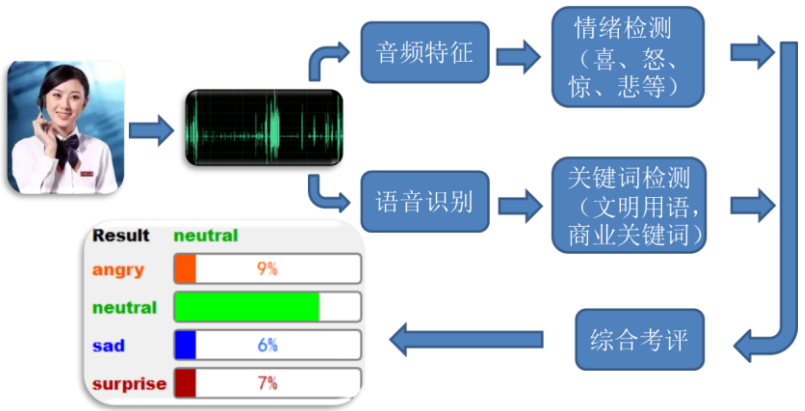
研究背景

- 多种应用
  - 智慧教学
  - 智慧营销
  - 客服质检
  - 案件侦破
  - ……
- 价值
  - 提升生活质量
  - 增加社会效益
  - ……



智慧教学

智慧营销



客服质检

案件侦破

方案设计

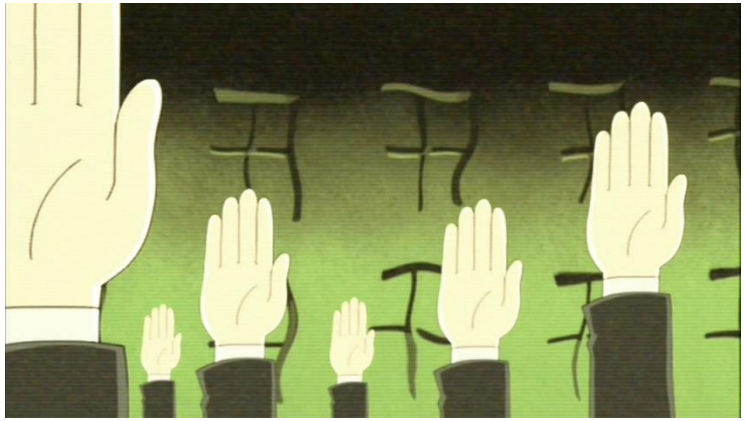
- 特征趋势

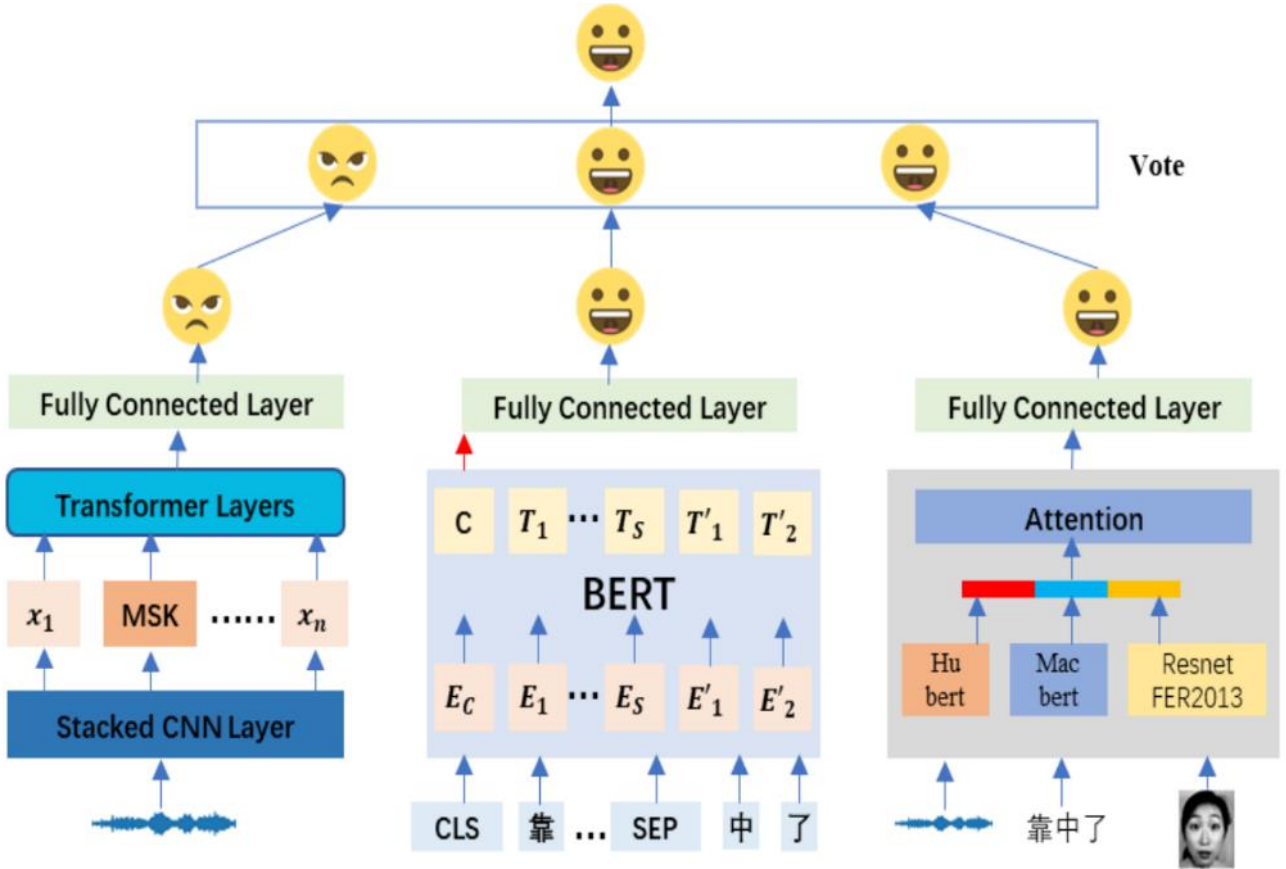
手工设计特征 → 深度学习模型特征 → 大规模数据预训练模型特征 → **微调预训练模型特征**

- 多模态融合

早期阶段（不同模态特征） → 中期阶段（不同模型） → **后期阶段（决策层级）**

- 模型结构设计
  - 投票





方案实现

模型1

基于Hubert（Hidden-Unit Bert）微调的语音模态情感识别模型

- 腾讯天籁实验室开源的中文语音自监督预训练模型HuBert
- 加一层全连接层，并冻结所有卷积神经网络层→用语音微调

维度为 1024

模型2

基于Macbert（MLM as correction BERT）微调的文本模态情感识别模型

- Bert基础上引入了纠错型掩码语言模型，缓解了“预训练-下游任务”不一致的问题
- Bert模型的pooling层后添加一层全连接层→用文本微调

维度为 1024

模型3

基于注意力机制融合的语音、文本和视频多模态融合的情感识别模型

- 预训练的Hubert、Macbert和Resnet-FER2013模型提取特征
- 句子级别特征拼接→基于注意力机制融合

Resnet-FER2013
维度为 512

数据集及预处理

- 中国人民大学提供的M^3ED多模态数据集

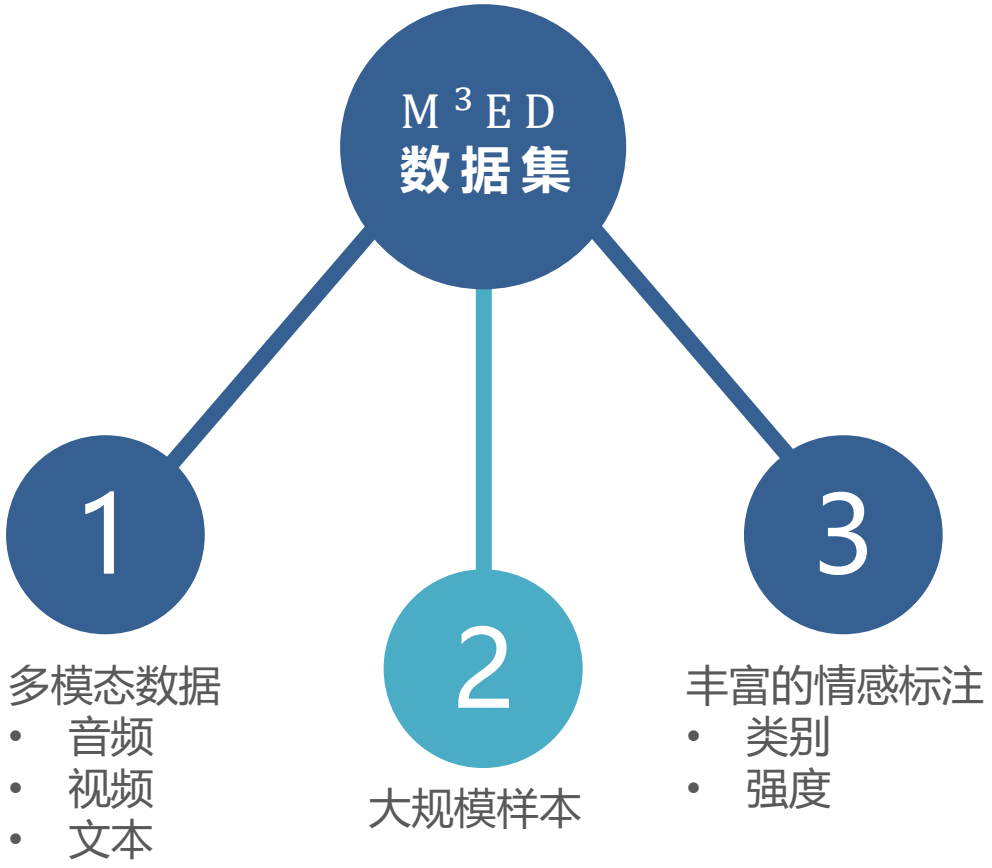
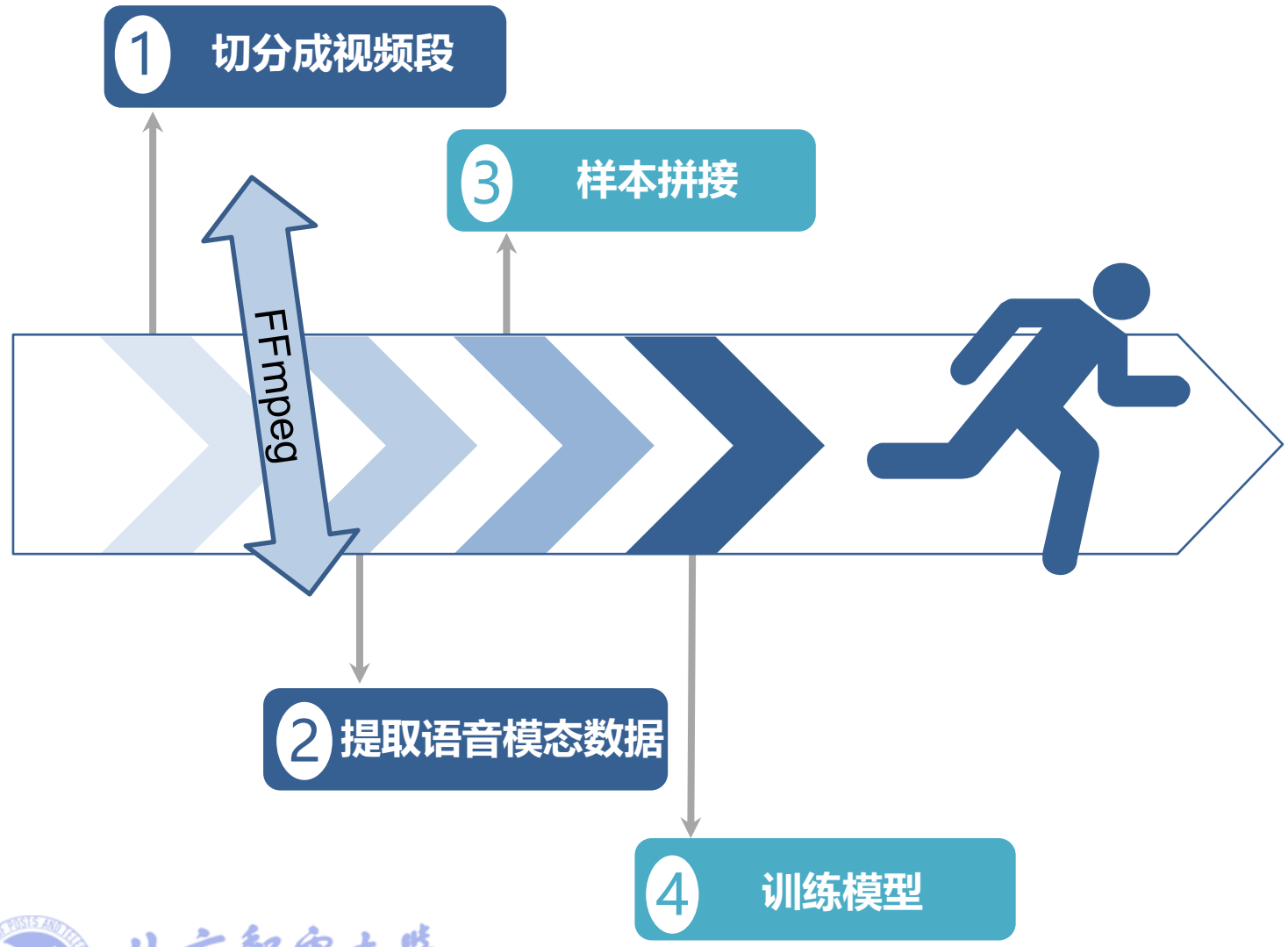


表1：M^3ED多模态数据集统计信息

	训练集	验证集	测试集
对话数量	685	126	-
对话轮数	6505	1016	1191
语句数量	17427	2821	-
说话人数量	421	87	-

数据集及预处理



算法1：数据样本拼接方法
输入：标注Json文件
输出：拼接完成的语音样本和文本
1: If 当前样本的说话人、话题编号和标签与 文件中下一条的样本一致 满足 do
2: 将音频名、文本和标签按顺序保存在序列中
3: Else
4: 加载现存的序列，读取音频和文件进行拼接，将多个样本拼接为一个样本。保存格式为：说话人_剧名_话题ID_1_2_3.wav(表示3个样本拼接在一起)
5: End

数据评估与总结

- 实验配置

子模型1和2

子模型3

学习率	1e-5	1e-4
优化器	AdamW	Adam
批样本数	1（4张卡，梯度累计步数为 2）	64（单卡）
训练Epoch数	100	
损失函数	交叉熵	

数据评估与总结

➢评估指标
- F1-score

➢讨论
- 模型1&2：语音>文本
- 多模态：
  验证集与1相近；
  测试集表现较差
- 投票策略最有效

➢总结
- 多模态+预训练微调+投票策略

	验证集	测试集
子模型1（语音）	0.6068	0.5127
子模型2（文本）	0.5192	-
子模型3（多模态）	0.6013	0.4937
投票集成	-	**0.5272**

情感识别小伙伴们

李启飞(博一)

王栋(准博一)

王聪(准研一)

任一鸣(大三)

高迎明(讲师)

李雅(副教授)

谢谢关注！

敬请批评指正！