1 方法

1.1 特征提取

本文使用 M³ED 数据集[1]，对多模态情感对话进行分析和识别。M³ED 数据集包含语音、图像和文本三种模态的信息，以及人工标注的情感标签。本文针对这三个模态，对每种模态都进行了特征提取。

为了提取语音模态的信息，本文首先使用 ffmpeg 工具，从视频文件中分离出音频信号，并保存为 wav 格式。然后，本文采用 VGGISH 模型提取音频特征。具体做法是将音频信号切分为多个帧，并对每帧进行变换，从中提取出 1024 维的音频特征。VGGISH 模型是一个基于 VGG 网络结构的深度神经网络，能够从原始音频信号中提取语音特征。由于 VGGISH 模型是在声音检测的数据集 AudioSet[2]上训练的，与情感识别任务不完全匹配。因此，本文使用 IEMOCAP 数据集[3]对 VGGISH 模型进行微调，使其能够适应不同的情感类别或维度。IEMOCAP 数据集是一个多模态情感对话数据集，包含了 10 个说话人的 12 小时的对话录音和情感标签。最后，本文在 M³ED 数据集上使用微调后的模型提取音频特征，并对每个句子对应时间内的多帧音频特征进行平均，以得到一个句子级别的语音特征，用于情感预测。

为了提取图像模态的信息，本文首先使用 openface 工具，从视频中检测和裁剪人脸图像，使用类似 TAE 模型[4]的方式来提取图像特征。TAE 模型是一种从无标注人脸视频中学习面部动作表征的方法，它认为两帧连续的人脸图像（源和目标）之间的变化包括了面部动作和头部姿态的信息。因此，TAE 通过建模这种信息用来描述面部表情特征。本文对每个句子对应时间内的多张图像特征进行平均，得到一个句子级别的图像特征，用于情感预测。

对于文本模态，本文使用 RoBERTa for chinese 模型[5]提取文本特征。RoBERTa for chinese 模型使用了 30G 的中文文本，包含 3 亿个句子，100 亿个字，进行了大规模的自监督学习，能够学习到丰富的语义信息。本文使用 RoBERTa for chinese 模型作为特征提取器，从文本输入中提取出 1024 维的语义特征向量，以适应 M³ED 数据集上的情感对话任务。

1.2 模型建立

本文分别从语音、图像和文本三个模态提取了 1024 维、256 维和 1024 维的特征，并通过归一化和两层 DNN 将其映射到 64 维的特征向量。然后，本文将三个模态的向量按顺序拼接成 192 维的模态特征向量。由于是双人对话，本文再将两个人的模态特征向量拼接成 384 维的输入特征向量。

为了充分利用双人谈话中的对话信息，本文实现了一个多模态特征拼接策略。该策略在预测每个人的情感时，不仅考虑说话人自身的视频、音频和文本特征，还考虑对话人的相应特征。如果对方没有说话，我们就用 0 填充对方的特征。这样可以在预测每个人的情感时，同时考虑他和对方的对话信息。如图 1 所示，在预测 A 的情感时：当 A 说话时，将 A 的三个模态的特征信息 f_a 提取出来。将 B 的说话信息当作补充信息 f_b，将 f_a 和 f_b 拼接起来作为 A 的对话信息。如果未检测到 B 说话，则将 A 的特征信息 f_a 和置为 0 的 B 的特征信息 f_b 拼接在一起。

为了充分利用上下文信息，本文在构建说话人的时序特征序列时，同时考虑了说话人和对话人的说话内容。本文将说话人和对话人的特征拼接后，按照时序顺序将上下文信息补充后排列成一个特征序列。如图 1 所示，在预测 B 的情感时：当 B 说话时，将 B 的三个模态的特征信息 f_b 提取出来，将 A 的特征作为补充与 B 的特征拼接在一起，如果未检测到 A 说话，则将 B 的特征 f_b 和置为 0 的 A 的特征 f_a 拼接在一起。除此之外，本文将上下文信息也添加进特征序列中，即当 B 未说话时，而 A 说话时。将置为 0 的 B 的特征 f_b 和 A 的特征 f_a 拼接在一起，并将其以时间顺序插入作为上下文信息。

最后，将得到的时序特征送入一个共享的两层双向 LSTM 中，分别输出说话者 A 和 B 的情感预测结果。

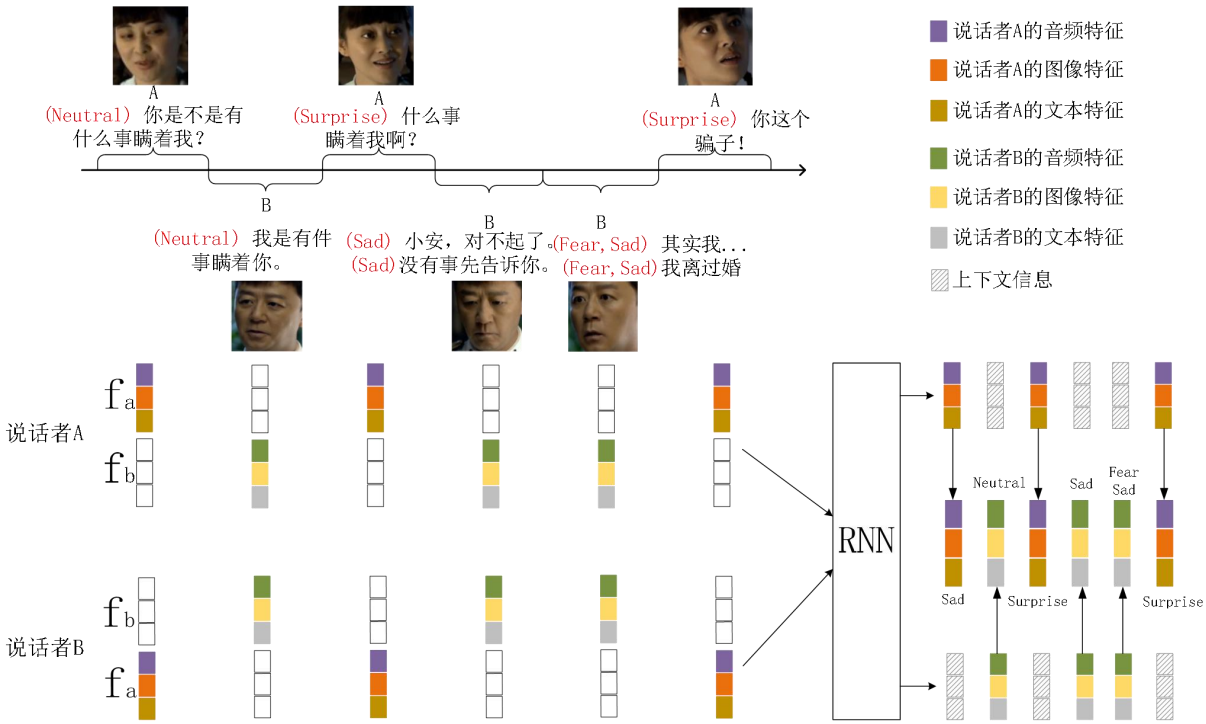

图 1 总体流程图

2 实验

2.1 数据集与评价指标介绍

M³ED 是一个中文的多模态情感对话数据库，收集了来自 56 部电视剧的 990 个双人对话，共有 9,082 个轮次和 24,449 个语句。M³ED 对每个语句进行了 7 种情感类别（快乐、惊讶、悲伤、厌恶、愤怒、恐惧和中性）的标注，并包含了语音、图像和文本三种模态的数据。M³ED 是一个具有跨文化价值的情感对话数据集，涵盖了多种影响情感的对话场景，并利用了声音、图像和文本三种模态的互补性。

本文采用 Macro-F1 作为评价指标。Macro-F1 指标是对不同类别的 F1 值求平均，能够体现模型在各个类别上的综合性能，而不会受到类别不平衡的影响。

2.2 数据处理细节

为了将图像模态和音频模态在时间位置上对齐，本文将所有视频处理为 25fps,在提取音频特征时，对应设置窗长为 100ms，窗移为 40ms。为了处理一个 batch 中视频长度不同的问题，本文采用了补齐策略使所有视频长度一致，并在计算损失函数时只考虑有效长度。

2.3 实验参数设置

本文设置 batchsize 为 8，epoch 为 15，本文采用两层双向 LSTM 结构，其隐藏层的维度设置为 128。同时，我们只使用有效长度进行训练，并采用加权交叉熵作为损失函数，使用 Macro-F1 值作为评估指标。

2.4 实验结果

本文以 Macro-F1 为指标，从验证集上选择了最优的模型在测试集上进行测试。表 1 展示了在 M³ED 数据集上进行两次实验的结果，实验 1 中，设置学习率为 3e-4,此时训练集的识别准确率为 0.76，验证集上的准确率为 0.665，Macro-F1 值为 0.562，测试集上的 Macro-F1 值为 0.565；实验 2 中调整学习率为 1e-4，此时在训练集、验证集以及测试集上的识别结果略有提升。

表 1 多次实验结果展示

实验	训练集	验证集		测试集
	(ACC)	ACC	Macro-F1	(Macro-F1)
实验 1	0.76	0.655	0.562	0.565
实验 2	0.78	0.70	0.57	0.566

参考文献

[1] Zhao J, Zhang T, Hu J, et al. M3ED: Multi-modal multi-scene multi-label emotional dialogue database[J]. arXiv preprint arXiv:2205.10237, 2022.

[2] Gemmeke J F, Ellis D P W, Freedman D, et al. Audio set: An ontology and human-labeled dataset for audio events[C]//2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2017: 776-780.

[3] Busso C, Bulut M, Lee C C, et al. IEMOCAP: Interactive emotional dyadic motion capture database[J]. Language resources and evaluation,

2008, 42: 335-359.
[4] Li Y , Zeng J , Shan S .Learning Representations for Facial Actions from Unlabeled Videos[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2020, PP(99):1-1.
[5] Liu Y , Ott M , Goyal N ,et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach[J].2019.