陕西科技大学
SHAANXI UNIVERSITY OF SCIENCE & TECHNOLOGY
电子信息与人工智能学院
School of Electronic Information and Artificial Intelligence

基于时序上下文交互信息联合训练的多模态对话情感识别

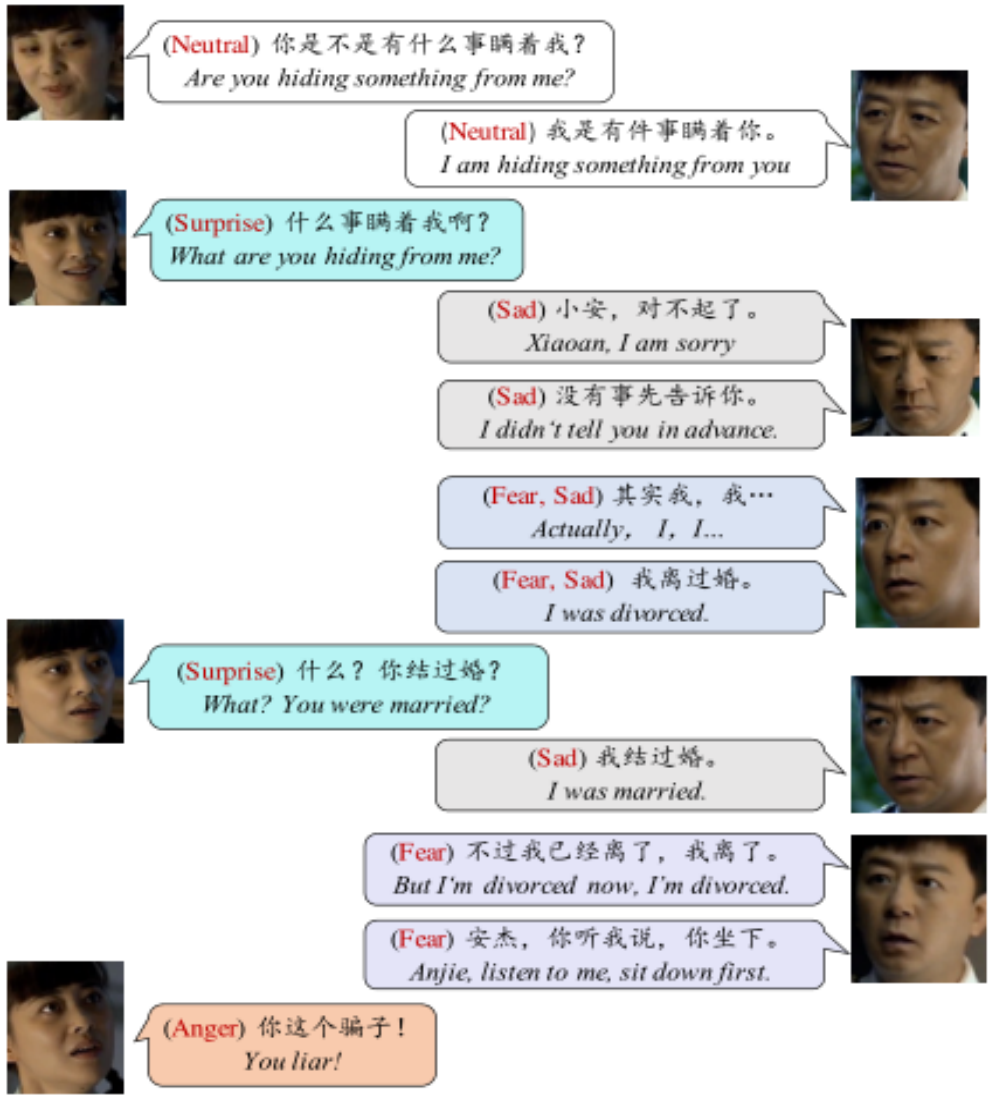
报告人： 李晶

队伍名称：SUST-EiAi-TEAM

队伍成员：陈海丰、李晶、张倩、郭楚佳、白义民、陈景霞

目 录
CONTENTS

一、 数据集介绍



- M^{3}ED[1]是首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。
- M^{3}ED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。
- M^{3}ED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。

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

二、 多模态特征提取

语音特征	图像特征	文本特征

1. **使用ffmpeg工具从M^{3}ED数据集中分离出语音信号**
2. **使用IEMOCAP数据集[1]对VGGISH模型[2]进行微调**
3. **使用微调后的VGGISH模型提取语音特征**
4. **得到句子级别的语音特征**

[1] 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.
[2] Hershey, Shawn , et al. "CNN architectures for large-scale audio classification." IEEE International Conference on Acoustics IEEE, 2017.

二、 多模态特征提取

语音特征	图像特征	文本特征

1. **使用openface工具检测和对齐人脸图像**
2. **使用我们最新研究的一种自监督面部动作表征学习方法[1] 来提取面部动作特征**
3. **得到句子级别的图像特征**

[1] self-supervised facial action unit representation learning with ensembled prior constraints. under review

二、 多模态特征提取

语音特征	图像特征	文本特征

1. 使用RoBERTa for chinese模型[1]提取文本特征

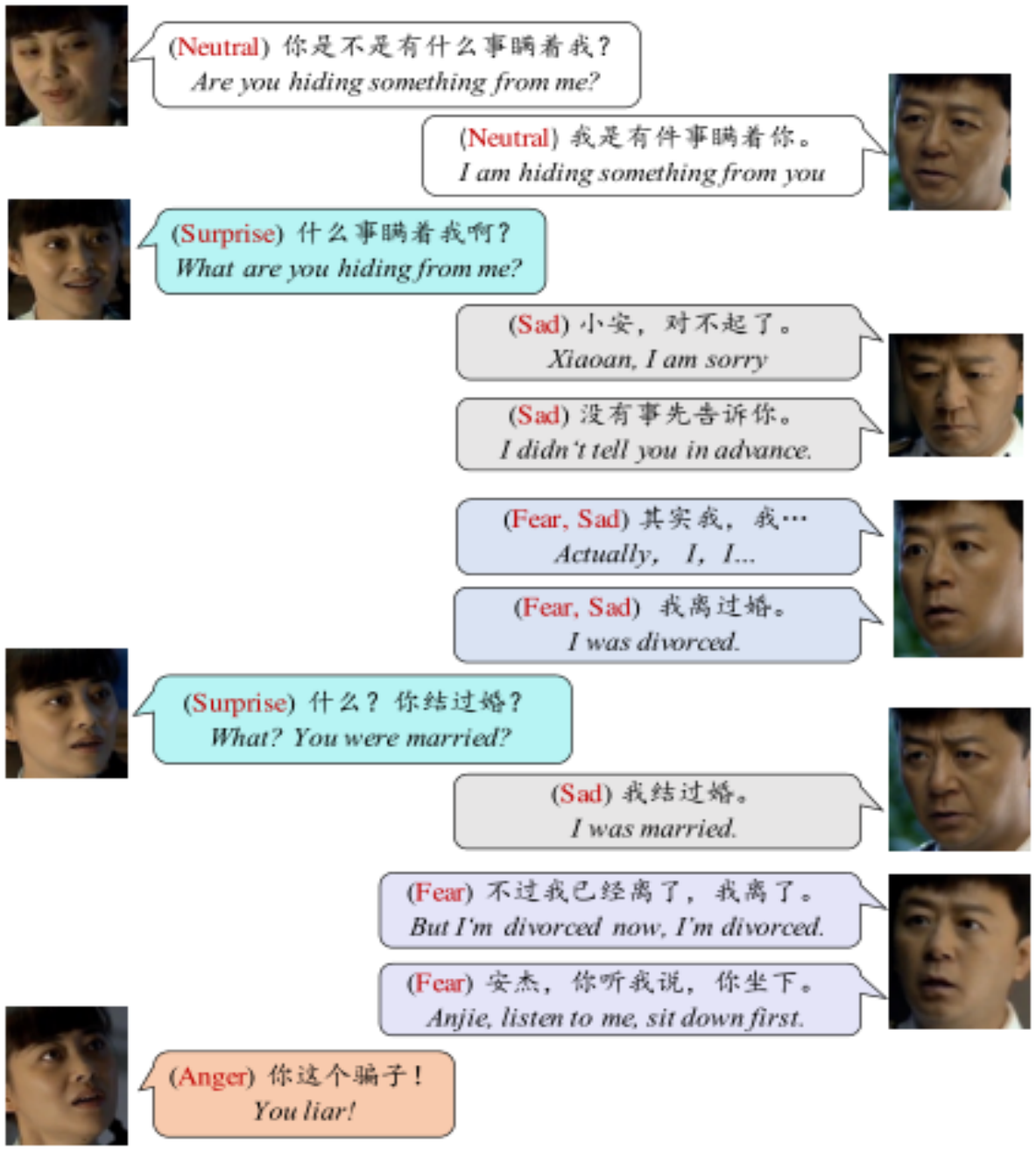
RoBERTa for chinese模型使用了30G的中文文本，包含3亿个句子，100亿个字，进行了大规模的自监督学习。

2. 得到句子级别的文本特征

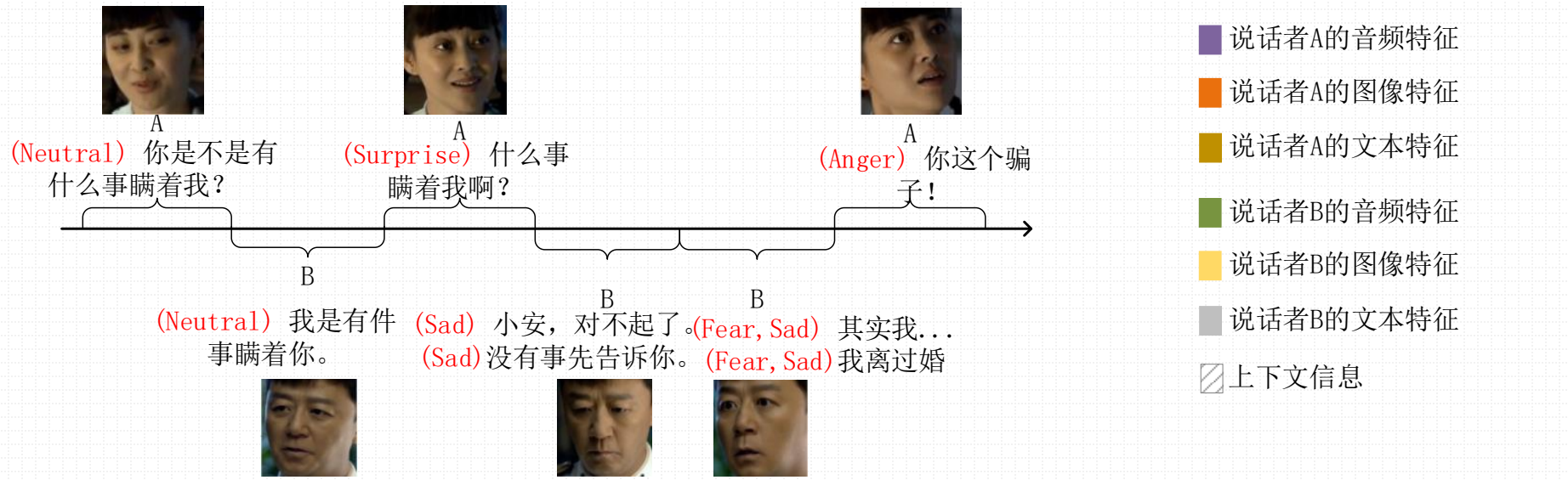
[1] Liu Y , Ott M , Goyal N ,et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach[J].2019.

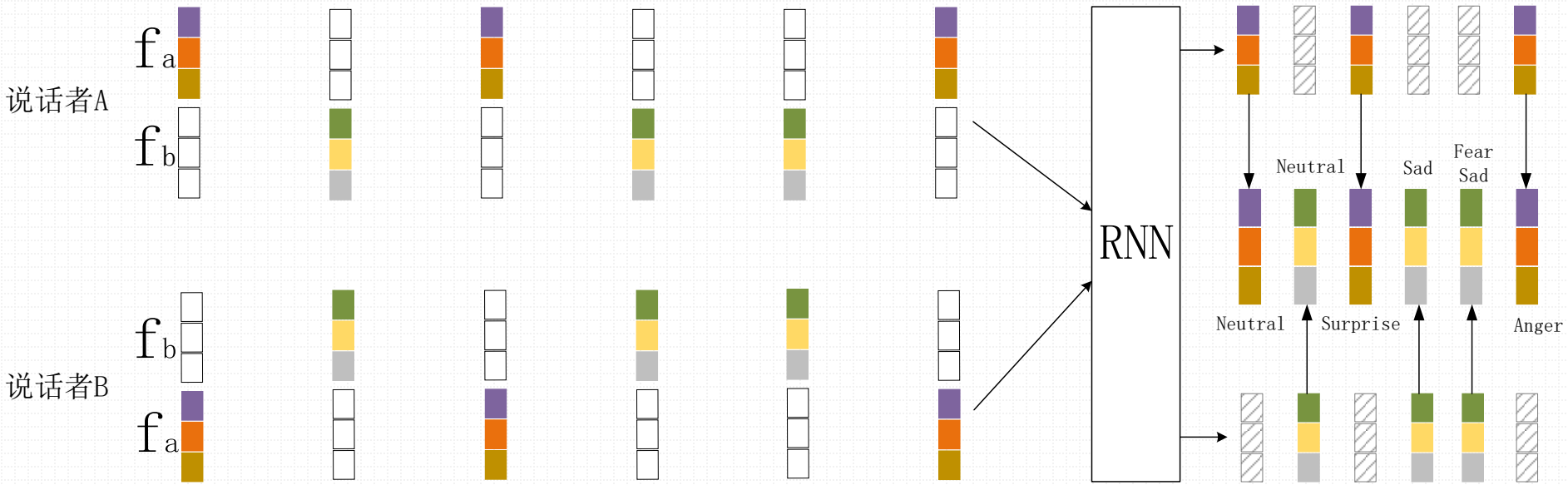
三、 模型介绍

- 多模态融合
- 交互信息
- 上下文交互信息
- 联合训练



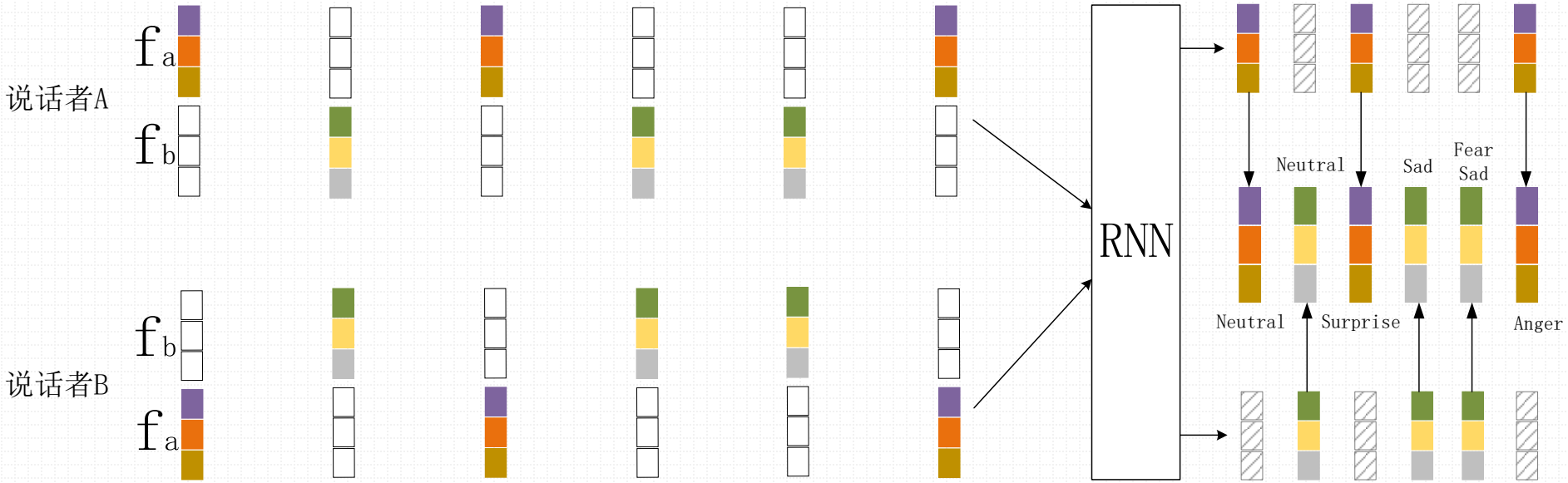
三、 模型介绍





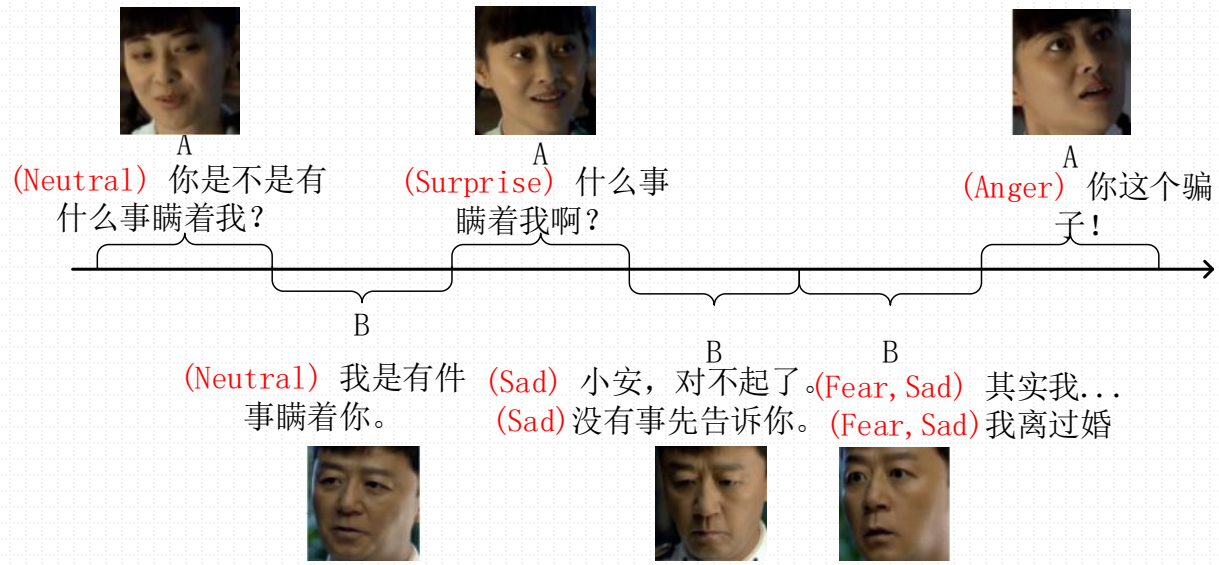
- **多模态融合**
- 将1024维语音特征、256维图像特征和1024维的文本特征，并通过归一化和两层DNN将其映射到64维的特征向量。
- 将三个模态的特征按顺序拼接成192维的特征向量。

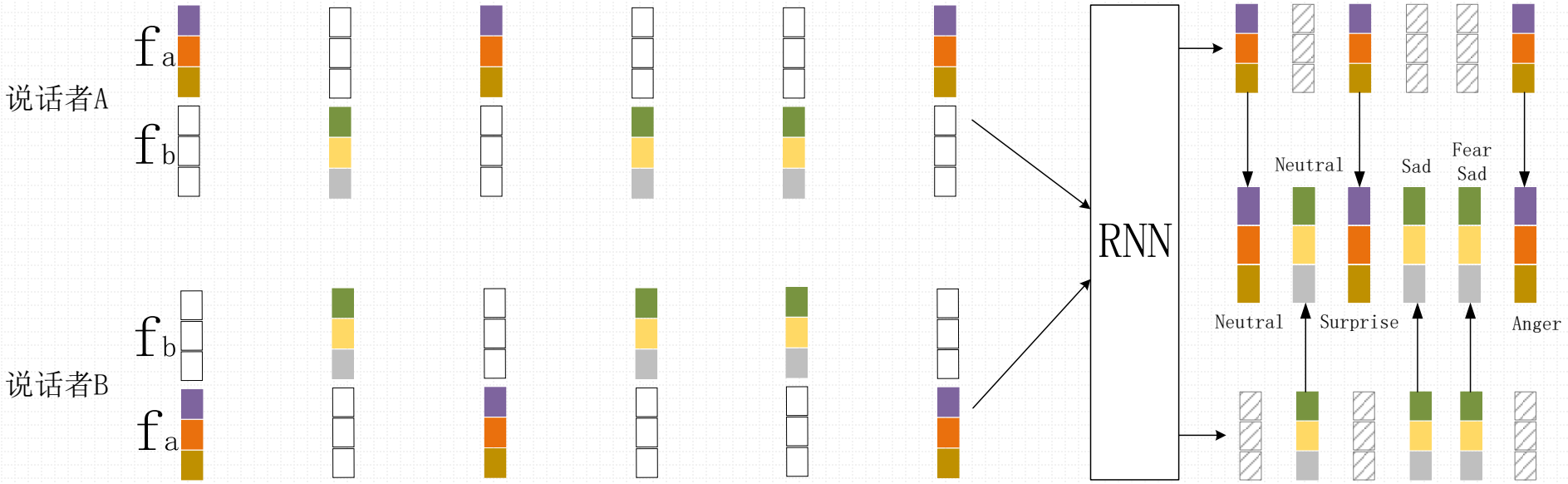
三、 模型介绍



- **交互信息**
- 对话中不同说话者之间会相互影响，例如一个说话者的情感可能会受到另一个说话者的话语的刺激。
- 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。

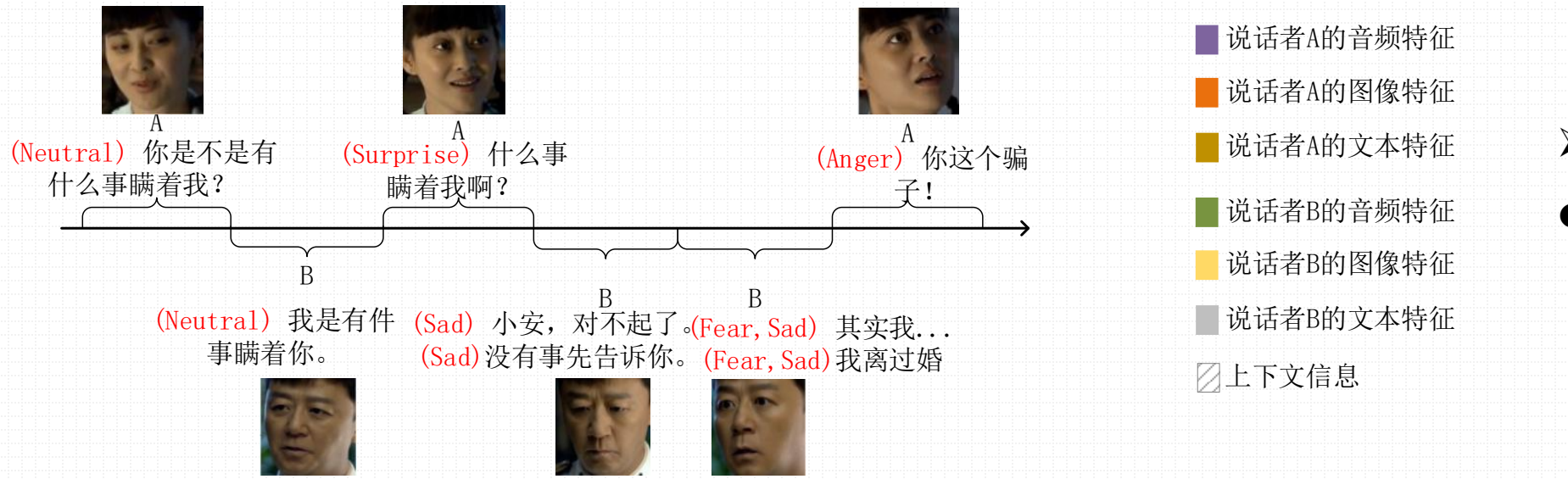
三、 模型介绍

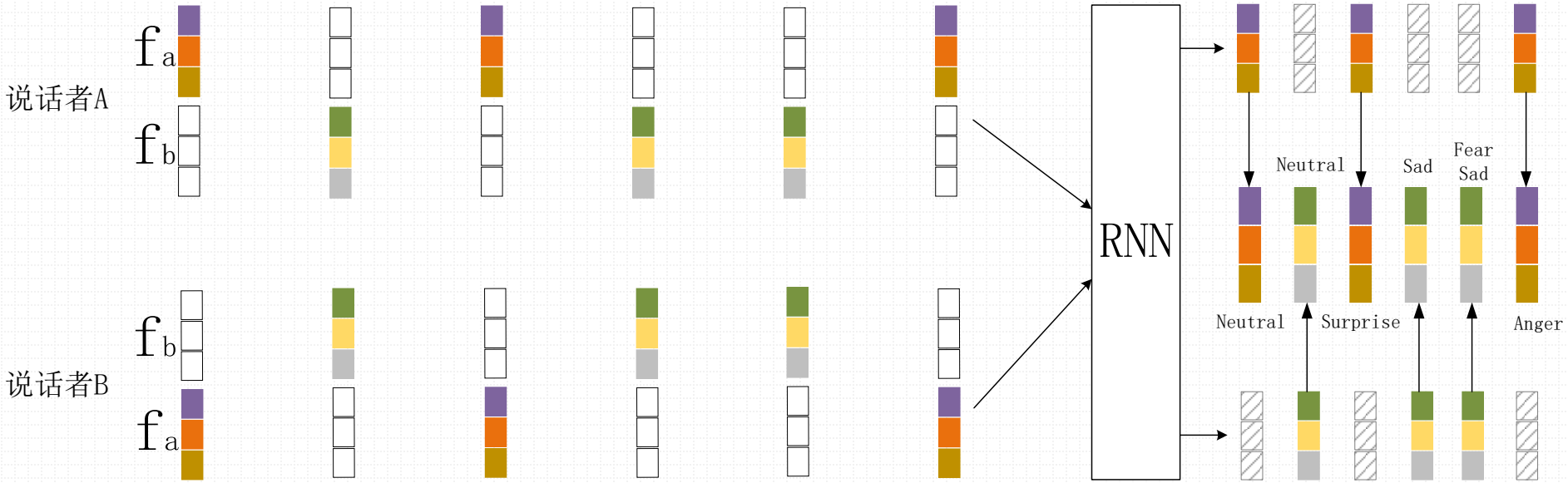




- **时序上下文交互信息**
- 一个说话者的情感可能会随着对话的进展而发生转折或持续。
- 在构建说话人的特征序列时，将对话人的融合特征以时序方式插入特征序列中，作为时序上下文交互信息。

三、 模型介绍





- **联合训练**
- 将得到的时序特征序列送入共享的两层双向LSTM中，编码得到了说话和对话者相应的包含时序上下文交互信息的高级表征。
- 使用说话人A和说话人B在对话时的特征，分别预测说话人A和说话人B的情感。

四、 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。下表展示了在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。

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

感谢聆听！