Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

一段话，多个情绪？模型如何识别“情绪变化”的蛛丝马迹

硕士研究生 杨桢弘

2025年04月13日

问题回溯

- 总结反思
  - 读PPT内容占比大，缺少互动
  - 算法部分讲解不够深入
- 相关内容
  - 2024.11.10 杨桢弘《深度学习语音情绪识别技术》

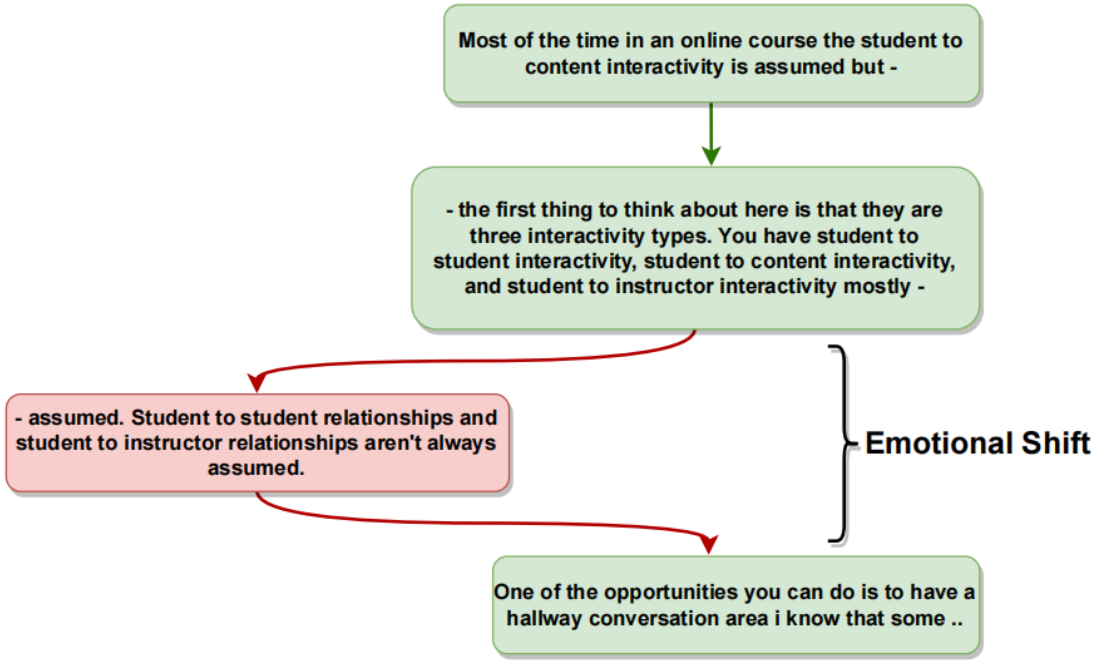
内容提要

- 预期收获
- 内容引入
- 内涵解析与研究目标
- 研究背景与意义
- 研究历史与现状
- 知识基础
- 算法原理
  - PCGNet
  - ENT/FENT
- 特点总结与未来展望
- 参考文献

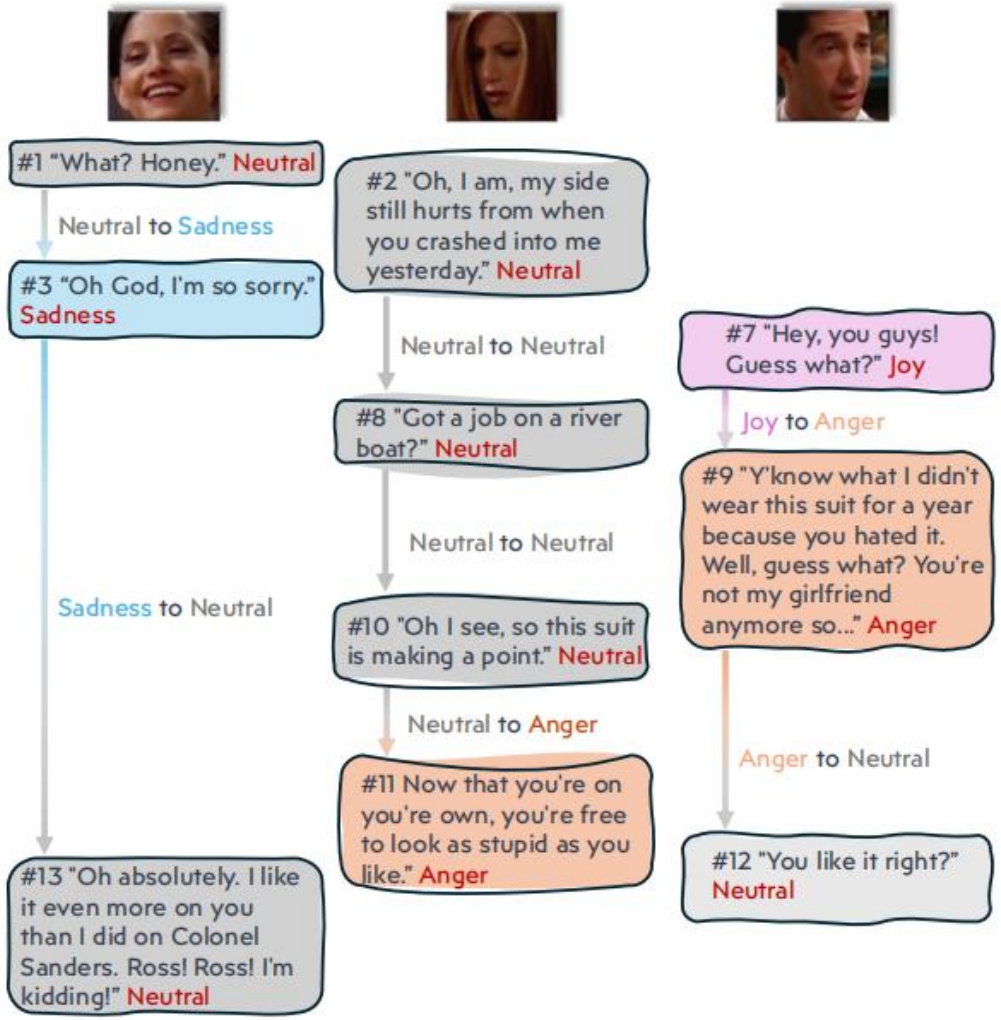
预期收获

- 预期收获
  - 了解对话中情绪变化识别和语音情绪识别的区别与联系
  - 掌握对话中情绪变化识别的基本模型及其原理
  - 明确对话中情绪变化识别未来发展方向

内容引入



neutral	***Phoebe***: And you can just tell me about yourself.
neutral	***Jim***: All right.
neutral	***Phoebe***: Okay.
neutral	***Jim***: I write erotic novels, for children.
surprise	***Phoebe***: What?!
neutral	***Jim***: They're wildly unpopular.
disgust	***Phoebe***: Oh my god.
neutral	***Jim***: Oh also, you might be interested to know that I have a Ph.D.
surprise	***Phoebe***: Oh my god. Wow! You do?
joy	***Jim***: Yeah, a Pretty Huge.
disgust	***Phoebe***: All right.



对话过程中的情绪变化

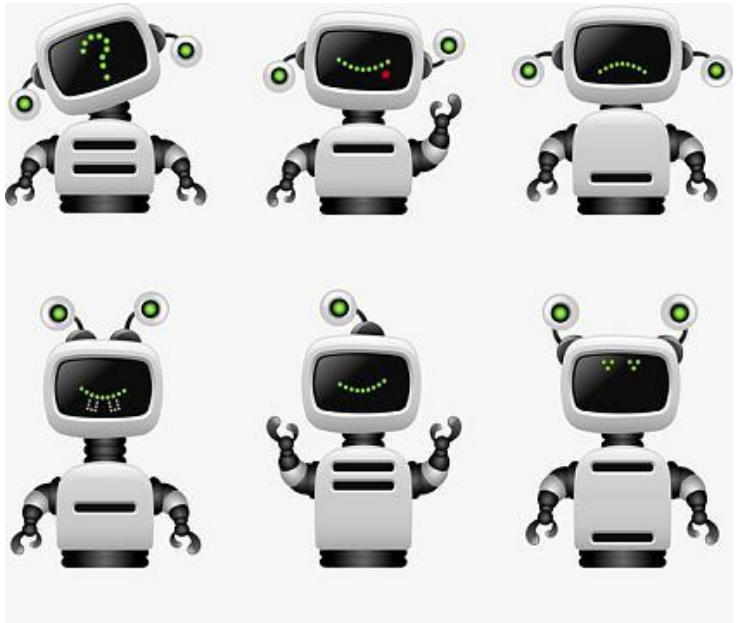
内涵解析与研究目标

- 内涵解析
  - 语音情绪识别
    - 语音→情绪类别/强度
    - 侧重单句分析，更关注情绪的分类
  - 对话中情绪变化识别
    - 语音/文本→情绪变化序列/情绪对
    - 侧重整个对话的分析，关注情绪的变化

- 研究目标
  - 结合深度学习、信号处理等技术
  - 在对话中识别情绪类别和强度的同时注重情绪变化，能够准确、高效的识别出情绪类别，并在情绪变化时及时做出改变

研究背景与意义

- 研究背景
  - 传统语音情绪识别（SER）大多关注单句级别的情绪分类，难以捕捉细粒度的情绪转移和上下文依赖
  - 随着Transformer、图神经网络（如DialogueGCN）以及跨模态建模方法（如PCGNet）不断涌现，为建模对话中的情绪依赖关系提供了新路径
- 研究意义
  - 更符合真实对话场景，能够更准确识别对话中每轮甚至每一刻的情绪状态及其变化
    - 心理健康
    - 智能客服
    - 教育系统



研究历史与现状

Ghosal 等人提出了DialogueGCN，首次在ERC中引入图神经网络，显式建模对话结构+说话人间关系

Harsh等提出了结合情绪突变检测模块(Emotion Shift Component)的多模态对话情绪识别模型，能够在检测情绪的同时输出情绪变化概率

Chen等人提出了一种新的结构DST(Deformable Speech Transformer)，通过动态窗口+偏移机制自适应聚焦有情绪价值的信息

Geng等人提出了PCGNet框架，将对话情绪识别与情绪变化检测任务统一建模，融合人格引导的交互建图与对比学习机制，显著提升模型对真实对话中情绪波动的感知能力

2021

Jiao等人首次明确建模情绪变化趋势与相互影响，构建了情绪迁移路径图，使用结构感知情绪演化路径，支持情绪转换检测任务

2022

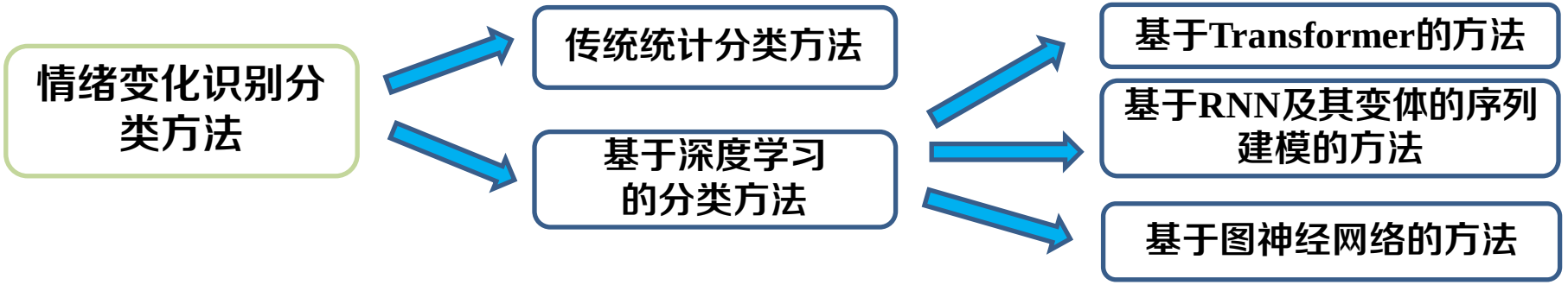
Anwer等人聚焦于从连续语音中识别情绪类别的“变化点”与“变化序列”，输入连续语音序列（非预切分），检测情绪变化点

2023

Chen 等人提出了 Speech Former++，遵循语音的结构性，使用分层Transformer + 高效编码方式，构造帧级、音素级、词级、句子级处理流程

2024

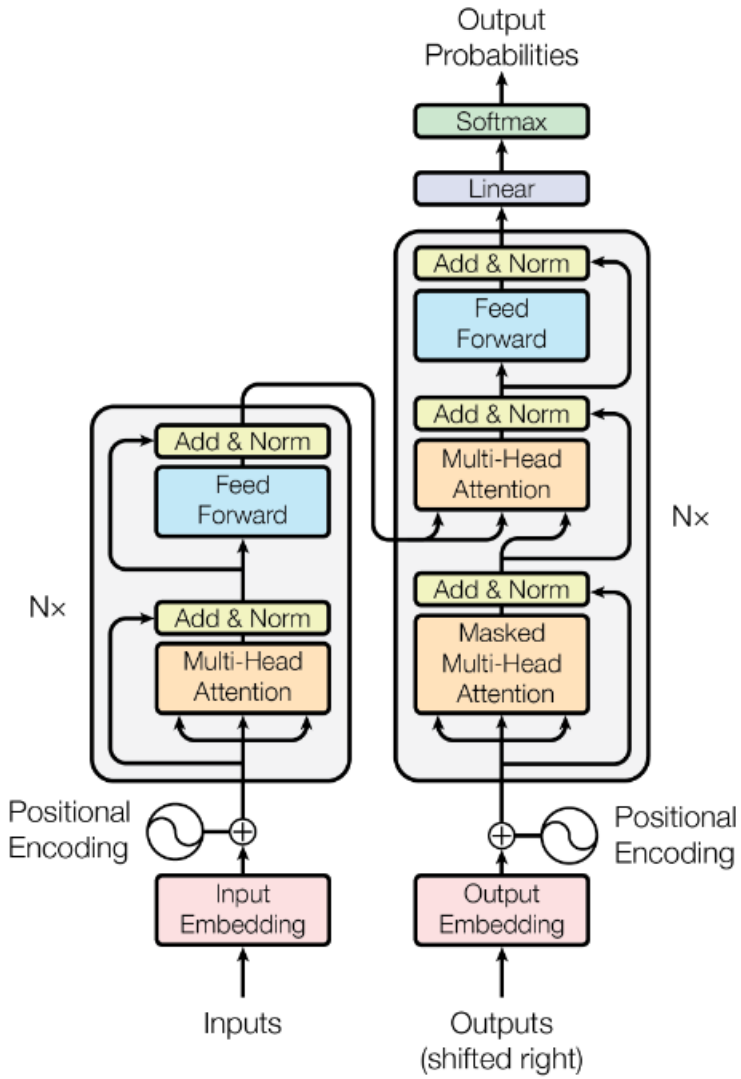
Li等人提出Emotion Neural Transducer框架，通过引入情绪对齐网格图和最大激活帧等机制，实现了对语音中细粒度情绪变化过程的高精度识别



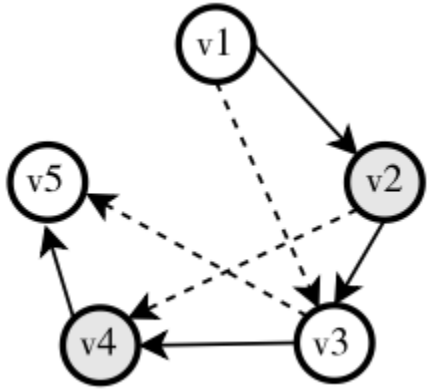
研究历史与现状

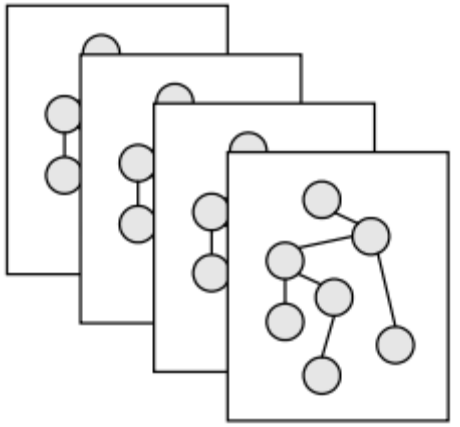
- 现有存在问题
  - 带标签的数据集少，且质量不高
  - 同一句话可能蕴含多种情绪，变化的边界模糊
  - 跨说话人建模、模态对齐、时序依赖建模难度高
- 现有的常见方法
  - 基于Transfomer的方法
    - 全局注意力强、支持预训练、适用于多轮长对话
  - 基于RNN及其变体的序列建模的方法
    - 构建于序列模型之上，具备时间依赖建模能力；结构相对简单，计算效率较高，适用于中小规模数据
  - 基于图神经网络的方法
    - 利用图结构可建模跨轮、跨说话人的高阶情绪依赖关系；支持模态融合与多任务学习，便于扩展为情绪变化检测任务

- Transfomer架构
  - 输入表示（Embedding + 位置编码）
    - 每个 token 首先经过嵌入，然后加入位置编码，弥补模型无序建模的问题
  - 多头自注意力机制（Multi-head Self-Attention）
    - 输入序列的每个位置都能“关注”其他所有位置
    - 建立词与词之间的依赖关系
  - 残差连接和层归一化
    - 保持梯度稳定
  - 前馈全连接网络
    - 对每个位置独立使用两层线性变换 + 激活函数



- 图神经网络
  - 节点（Nodes）：
    - 每一个节点表示一个utterance（话语），也可以拓展为：[utterance + speaker] 组合
      [utterance + modality]（用于多模态）
  - 边（Edges）：
    - 说话人边：同一个说话人说的句子之间连边（intra-speaker）
    - 上下文边：相邻话语之间连边，表示对话上下文
    - 情绪传播边：可根据标注数据（shift方向）构造辅助边（如情绪上升/下降）
    - 多模态边：文本、语音节点间连接（用于模态融合）





- 情绪变化识别的评价指标
  - 从识别的角度采用不同的衡量指标
    - 预测准确率ACC
    - 综合评估精确率与召回率F1值
    - 词错误率WER

$$WER = \frac{S + D + I}{N}$$

  - S：替换（substitution）词数
  - D：删除（deletion）词数
  - I：插入（insertion）词数
  - N：文本的总词数

- 情绪变化识别的评价指标
  - 从变化的角度采用的衡量指标
    - 情绪对话错误率EDER

$$EDER = \frac{\text{插入} + \text{删除} + \text{替换错误数}}{\text{总数}}$$

使用最小编辑距离算法：模型预测的情绪序列变成真实标签序列所需的最少操作次数

- 例如：

参考情绪序列：[happy, sad, angry, neutral]

预测情绪序列：[happy, surprise, neutral, fear]

EDER=0.75

A Persona-Infused Cross-Task Graph Network for Multimodal Emotion Recognition with Emotion Shift Detection in Conversations

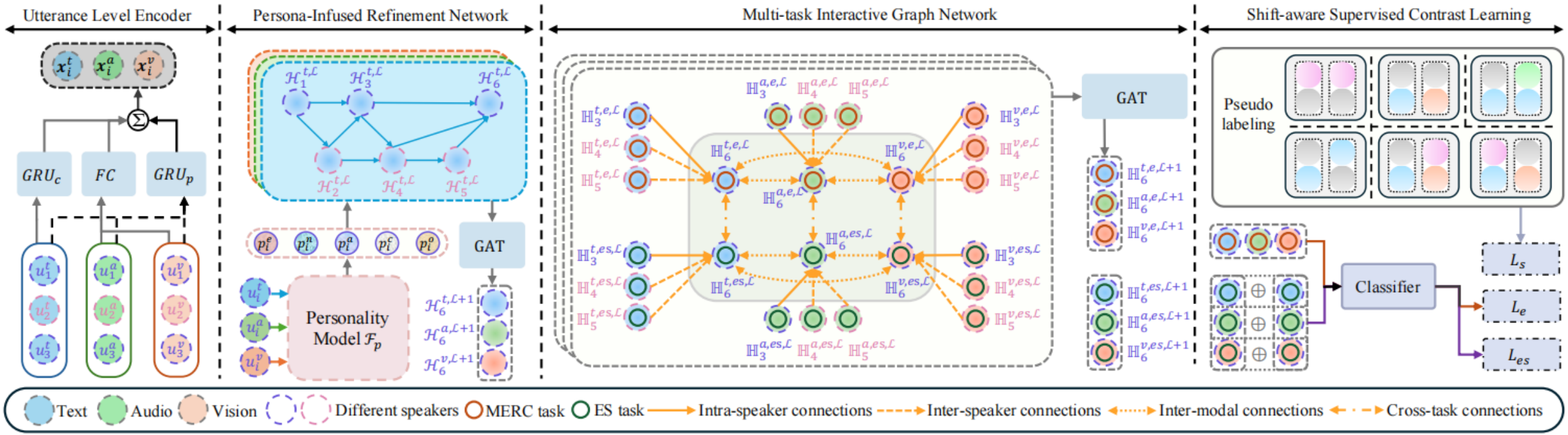
PCGNet TIPO

T	目标	提升上下文感知与人格因素建模能力，从而提高识别准确率
I	输入	2个数据集（IEMOCAP数据集：10位演员，151个对话，6类情绪；MELD数据集：13个角色，1433个会话，7类情绪）
P	处理	1. 构建人格感知图结构 2. 为 MERC 和 Emotion Shift Detection 构建图结构 3.情绪转换对比学习 4.预测情绪，识别情绪变化类型
O	输出	情绪序列（IEMOCAP:25种情绪转换；MELD: 36种情绪转换）

P	问题	1.现有方法忽略说话者与听者之间的互动差异 2.现有方法忽略情绪变化类型的检测
C	条件	需要单独的人格模型，需要预训练的Wav2vec2.0模型
D	难点	1.如何充分利用说话者个性特征 2.如何统一建模情绪分类与转变关系
L	水平	2024 CCF A类

PCGNet 算法原理

- 算法原理图
  - 构建人格感知图结构
  - 同时为 MERC 和 Emotion Shift Detection 构建图结构
  - 情绪转换对比学习



PCGNet 创新点分析

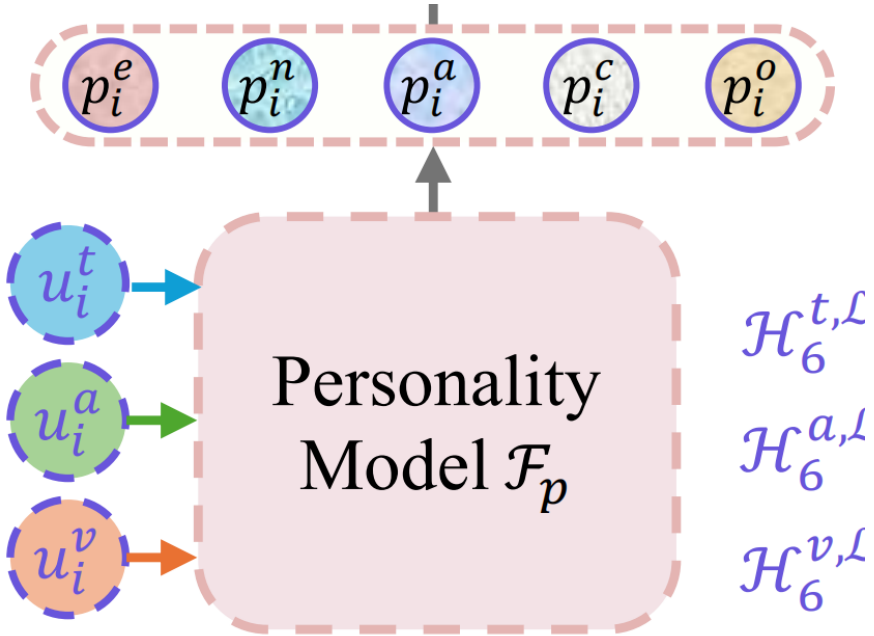
- 现有方法存在问题
  - 现有方法只建模说话人的身份，忽略说话者的人格对其说话的影响
  - 现有方法忽略说话者与听者之间的互动差异

- PCGNet的解决方法
  - 引入Big Five人格建模，构建一个“说话人与听者之间的人格交互图”，并用图注意力机制更新各个句子节点的表示，从而增强情绪表示的个性化与社交依赖性

- 解决方法 Persona-Infused Refinement Network
  - Big Five人格编码器
    - 对于每个说话人 s，提取其历史 utterance 序列，通过人格模型生成其人格向量 p
    - 生成的人格向量 p 会被用在Persona-Infused Refinement Graph中，构建边权重：

$$\Delta p_{ij} = |p_i - p_j|$$

    - p_i 、 p_j ： 说话人 i 和 听者 j 的人格向量
    - 越相似的人格 → 图中连接边越强 → 信息传播更顺畅
    - 动态分析不同情境下不同说话者表现出的大五人格特质



- 解决方法 Persona-Infused Refinement Network
  - Big Five人格编码器

维度缩写	含义	代表行为特征
O：Openness	开放性	有创造力、好奇心强、喜欢尝试新事物
C：Conscientiousness	尽责性	自律、组织能力强、做事可靠
E：Extraversion	外向性	善于社交、精力充沛
A：Agreeableness	宜人性	同理心强、乐于助人
N：Neuroticism	神经质	情绪不稳定、容易焦虑或烦躁

PCGNet 创新点分析

- 解决方法 Persona-Infused Refinement Network
  - Persona-aware 图神经网络
    - 每一个 utterance 表示 h_i 是一个图中的节点，图中包含：

      模态内连接边（如当前说话人前后 utterance 相连）

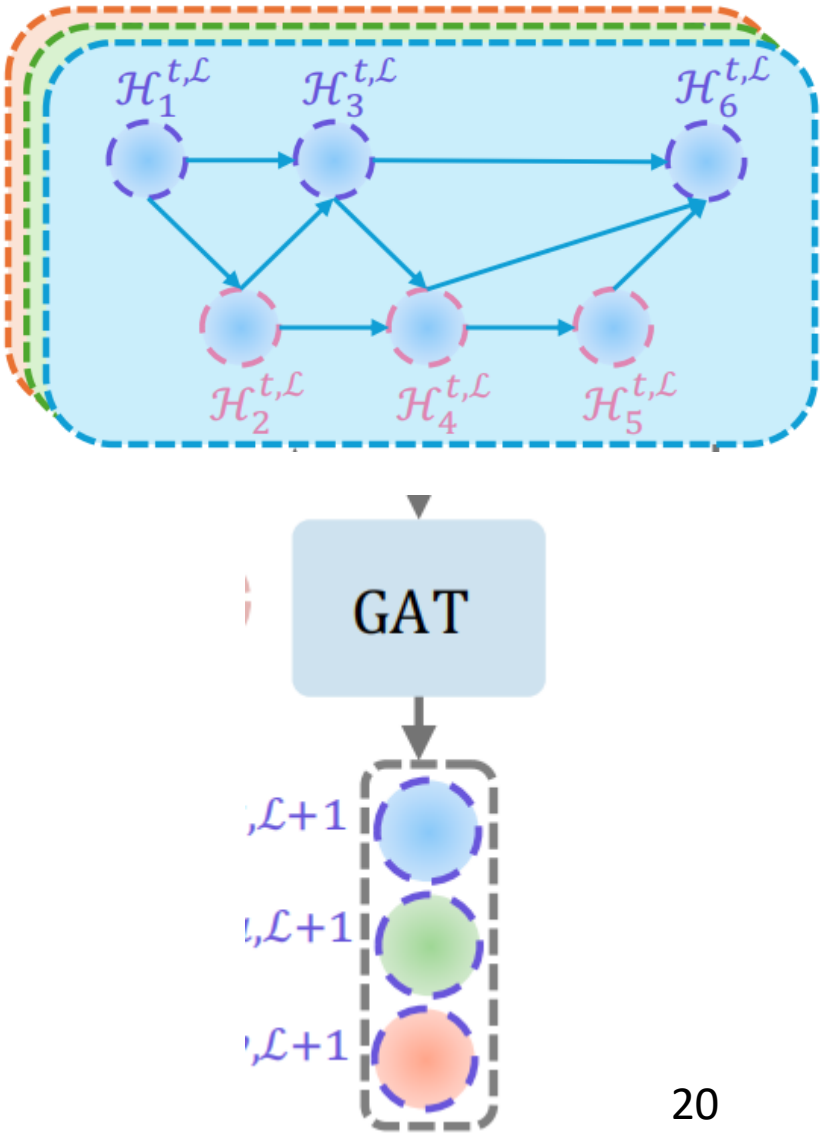
      模态间连接边（如同一句话的 text/audio/visual 相连）

      说话人–听者之间边（个性边）：

$$\alpha_{ij} = GAT(h_i, h_j, \Delta p_{ij})$$

$$\alpha_{i,j}^{\varepsilon,L-1} = Softmax(ReLU(a_p^T(W_p^{\varepsilon} p_i || W_p^{\varepsilon} p_j)))$$

      其中a_p^T为注意力参数， W_p^{ε}为人格特征权重矩阵
    - 图注意力聚合



PCGNet 创新点分析

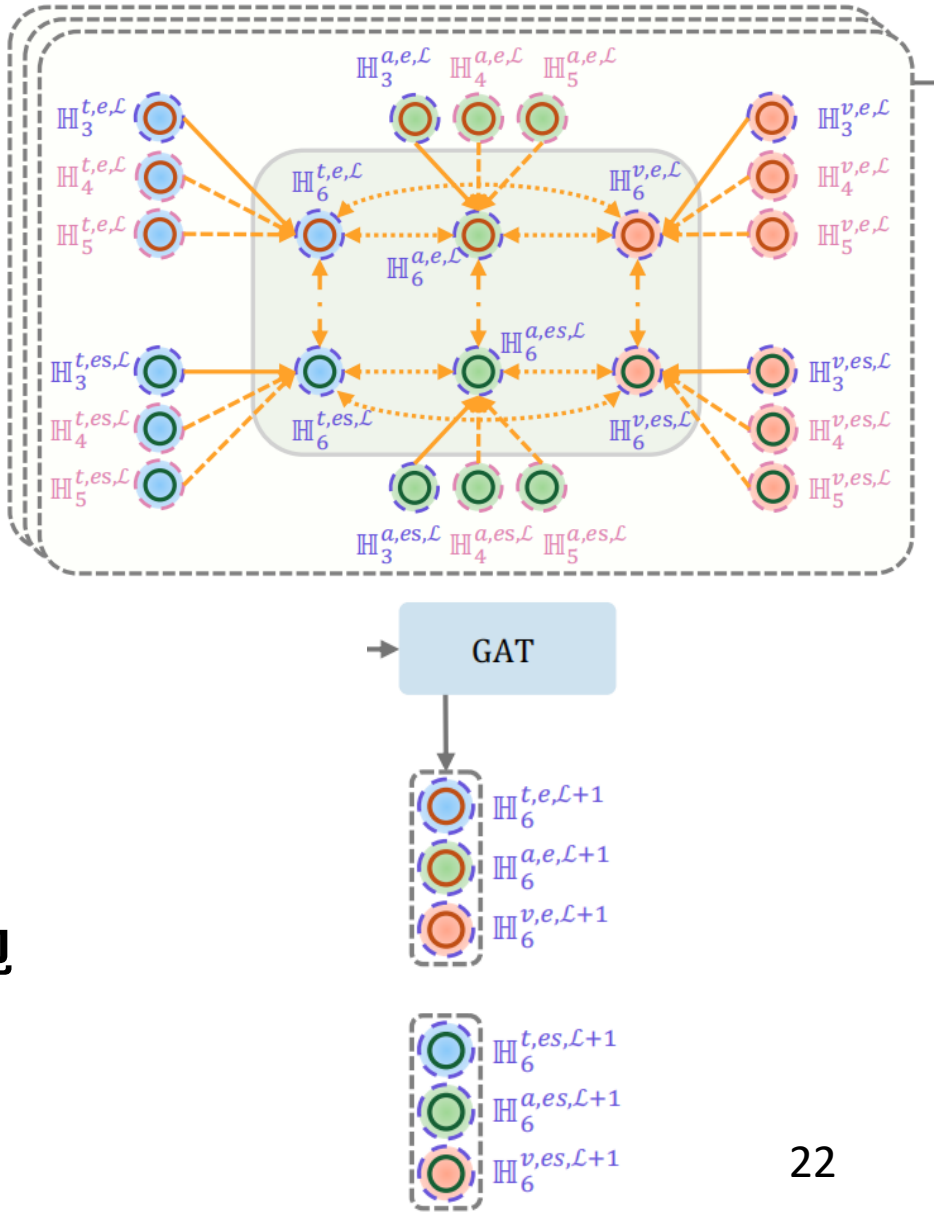
- 现有方法存在问题
  - 现有方法往往只建模主任务：对话中每句话的情绪分类（MERC），忽略了“是否发生了情绪变化（ES）”这一辅助任务中蕴含的信息
  - 现有方法大多只判断“是否发生情绪转变”，忽略了情绪变化的类型
- PCGNet的解决方法
  - 一个统一的图结构，将 MERC 和 ES Detection 两个任务联合建模，并允许它们相互传递信息
  - 通过对比学习方式，将相似的变化对聚在一起，区分不同变化模式

PCGNet 创新点分析

- 解决方法 Multi-task Interactive Graph Network
  - 节点（Nodes）
    - 对于每一个话语 u_i，我们为两个任务（MERC & ES）各建三个模态节点：

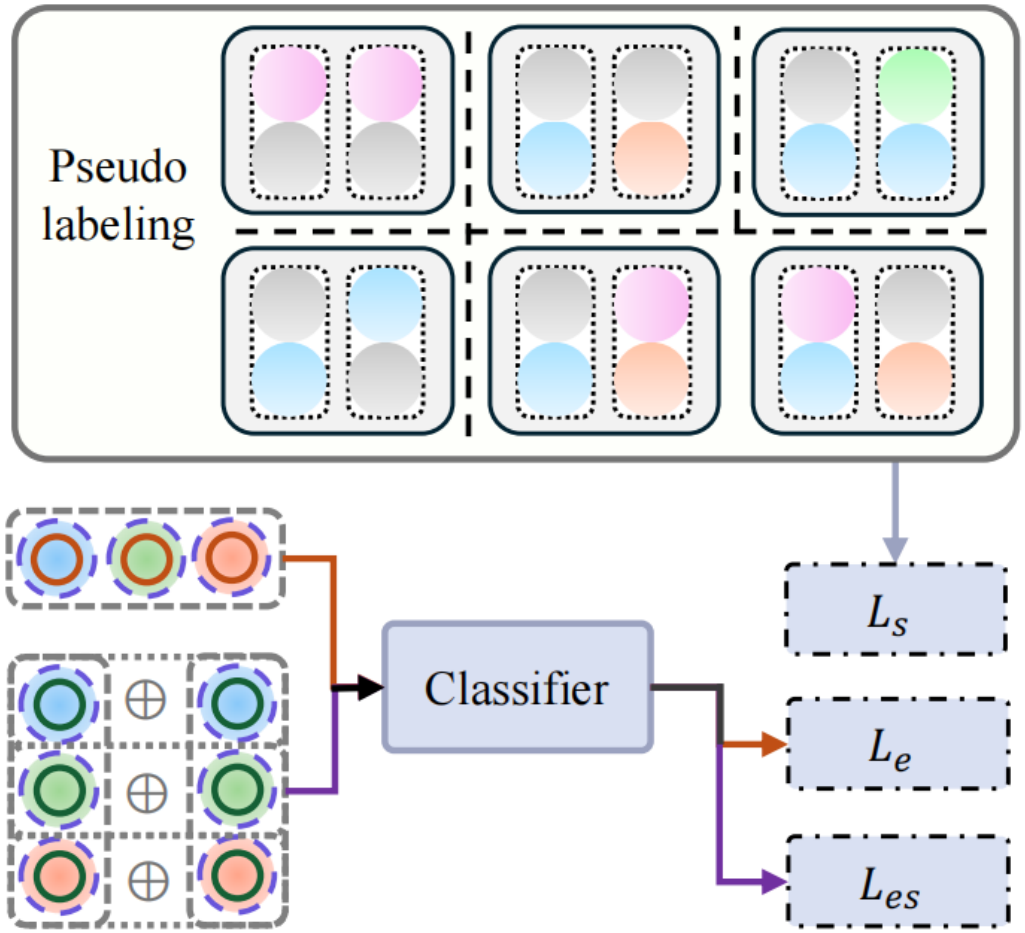
      MERC: $v_{t,e}$ $v_{a,e}$ $v_{v,e}$

      ES Detection: $v_{t,es}$ $v_{a,es}$ $v_{v,es}$
  - 边（Edges）
    - Intra-modal：同模态上下文关系，连接同一模态下的前后话语
    - Inter-modal：同一句话不同模态之间的连接
    - Cross-task：MERC 和 ES 的节点之间的连接（包括邻居）



PCGNet 创新点分析

- 解决方法 Shift-aware Supervised Contrast Learning
  - 输入：从 ES 模块中抽取连续话语对的多模态表示
  - 拼接：把前一句和后一句的表示合成“变化轨迹”
  - 分类：根据真实情绪标签对，将变化轨迹分成不同类别（伪标签）
  - 对比：比较每一对与其他对的相似度，训练模型拉近相同类，区分不同类
  - 目标：让模型学会“不同情绪变化类型”的特征差异，而不只是判断是否发生变化



实验设计　数据资源

- 数据集
  - IEMOCAP（10位演员，5男5女，151个对话，6类情绪，总时长11:28:12）
  - MELD（13个角色，1433个会话，7类情绪，总时长约13个小时）
- 对比方法

  DialogueRNN(2019)、DialogueGCN (2019)

  CTNet (2021)、MMGCN(2021)、MMDFN (2022)

  SCMM (2023)、CMCF-SRNet(2023)
- 评价指标
  - ACC：准确率
  - F1：加权F1分数

实验设计　对比实验

- 评估PCGNet在IEMOCAP和MELD数据集上的表现
  - PCGNet在IEMOCAP和MELD数据集上准确率和F1的平均值都优于对比方法，效果提升较大
  - 在MELD数据集上所有的类别都达到了最优，而IEMOCAP数据集上3类情绪是最优，其余也是次优

Methods	IEMOCAP								MELD								
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc	W-F1	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	Acc	W-F1
DialogueRNN♯	32.20	80.26	57.89	62.82	73.87	59.76	63.52	62.89	76.97	47.69	-	20.41	50.92	-	45.52	60.31	57.66
DialogueGCN♯	51.57	80.48	57.69	53.95	72.81	57.33	63.22	62.89	75.97	46.05	-	19.6	51.2	-	40.83	58.62	56.36
CTNet♮	51.30	79.90	65.80	67.20	**78.70**	58.80	68.00	67.50	77.40	52.70	10.0	32.50	56.00	11.2	44.60	62.00	60.50
MMGCN♯	45.14	77.16	64.36	68.82	74.71	61.40	66.36	66.26	76.33	48.15	-	26.74	53.02	-	46.09	60.42	58.31
MMDFN♯	42.22	78.98	66.42	69.77	75.56	66.33	68.21	68.18	77.76	50.69	-	22.93	54.78	-	47.82	62.49	59.46
SCMM♮	45.37	78.76	63.54	66.05	76.70	66.18	-	67.53	-	-	-	-	-	-	-	-	59.44
CMCF-SRNet♮	**52.20**	80.90	68.80	**70.30**	76.70	61.60	70.50	69.60	-	-	-	-	-	-	-		62.30
PCGNet(Ours)	49.83	**82.70**	**71.62**	69.14	76.08	**70.98**	**71.72**	**71.77**	**80.25**	**61.02**	**25.88**	**41.48**	**64.65**	**25.24**	**56.09**	**67.85**	**67.02**

实验设计　消融实验

- 评估PCGNet的不同模块在IEMOCAP和MELD数据集上的表现
  - w/o Persona-Infused：去除人格建模模块
  - w/o Shift-aware CL：去除对比学习模块
  - w/o Cross-task edges：去除跨任务图边

Methods	IEMOCAP				MELD			
	E-Acc	E-F1	ES-Acc	ES-F1	E-Acc	E-F1	ES-Acc	ES-F1
Ours	71.72	71.77	58.04	52.60	67.85	67.02	46.41	44.23
w/o Persona-Infused	70.86	70.73	57.78	52.04	66.78	65.99	45.33	43.11
w/o Shift-aware CL	70.92	70.96	56.76	51.27	67.05	66.13	45.82	42.91
w/o Cross-task edges	70.24	70.11	57.27	51.36	66.36	65.70	45.39	43.19

算法总结 PCGNet

- 算法贡献
  - Persona-Infused Refinement Network：解决了说话者与听者之间的互动存在差异的问题
  - Multi-task Interactive Graph Network：解决了情绪分类和情绪变化之间关系利用的问题
  - Shift-aware Supervised Contrast Learning：解决了情绪变化的类型和结构差异的问题
- 算法不足
  - 只关注了“变化对”，但并没有显式检测“变化点”
  - 人格维度（Big Five）是从模态特征“预测”出来的伪标签，可能存在不稳定或误导

ENT/FENT

Emotion Neural Transducer for Fine-Grained Speech Emotion Recognition

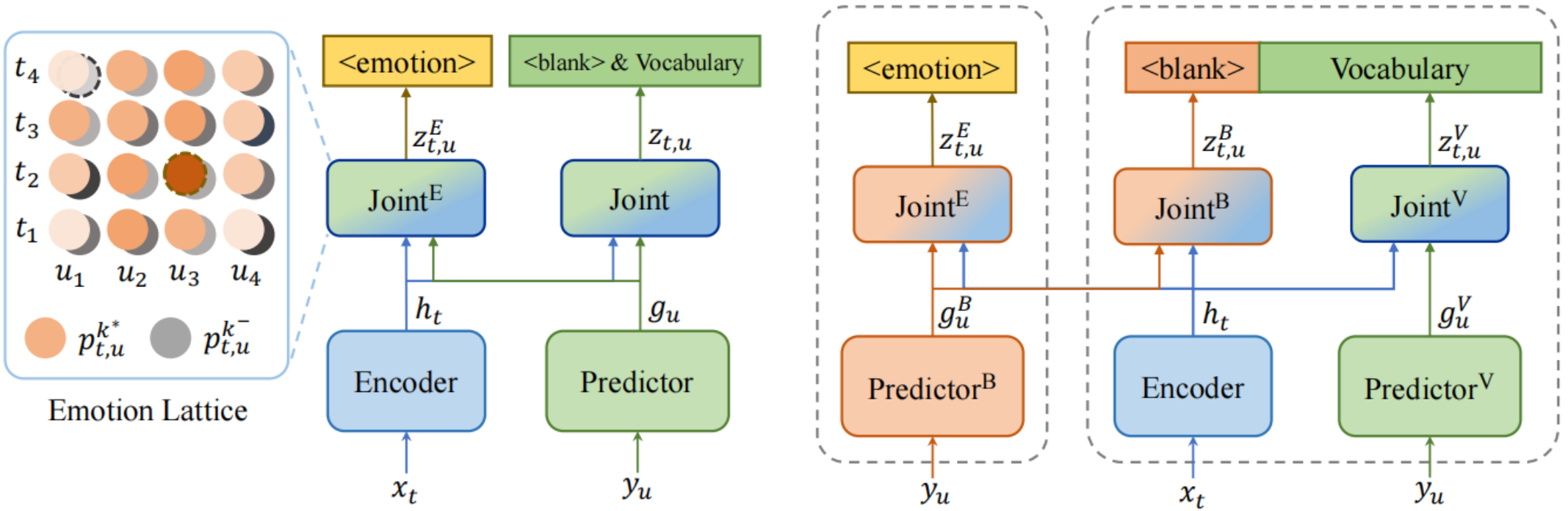
ENT/FENT TIPO

T	目标	实现帧级情绪识别与变化边界检测，提高语音情绪的时间建模精度
I	输入	2个数据集（IEMOCAP数据集：10位演员，151个对话，6类情绪；ZED数据集：73个说话者，180个语句，4类情绪）
P	处理	1.构建情感联合网络 2.设计格最大池化损失 3. 情绪边界建模 4.预测情绪的类别和发生转变的时间
O	输出	帧级情绪轨迹（IEMOCAP:25种情绪转换；ZED:9种情绪转换）

P	问题	1.现有方法忽略了单句语音中情绪变化的动态性 2.现有方法对非情绪帧和情绪帧区分不清
C	条件	需要预训练的Wav2vec2.0模型
D	难点	1.如何构建有效的帧×token对齐结构 2.如何在缺乏帧标签下学习时序情绪表示
L	水平	2024 CCF B类

ENT/FENT 算法原理

- 算法原理图
  - 构建情感联合网络
  - 设计格最大池化损失
  - 情绪边界建模



ENT/FENT 创新点分析

- 现有方法存在问题
  - 现有方法忽略了单句语音中情绪变化的动态性

- 解决方法
  - 情感联合网络（Emotion Joint Network）

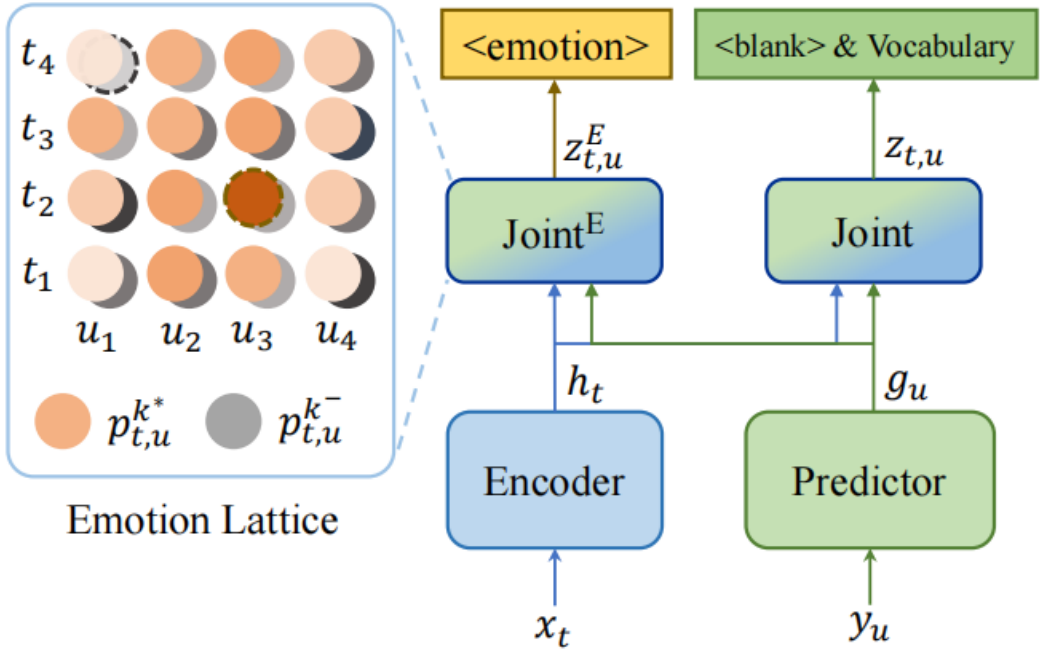
$$h_t = Encoder(x \le t)$$

$$g_u = Predictor(y \le u)$$

$$z^E_{t,u} = Joint^E(h_t, g_u)$$

$$p^k_{t,u} = softmax(z^E_{t,u})$$

其中$z^E_{t,u}$用来预测下一个输出



ENT/FENT 创新点分析

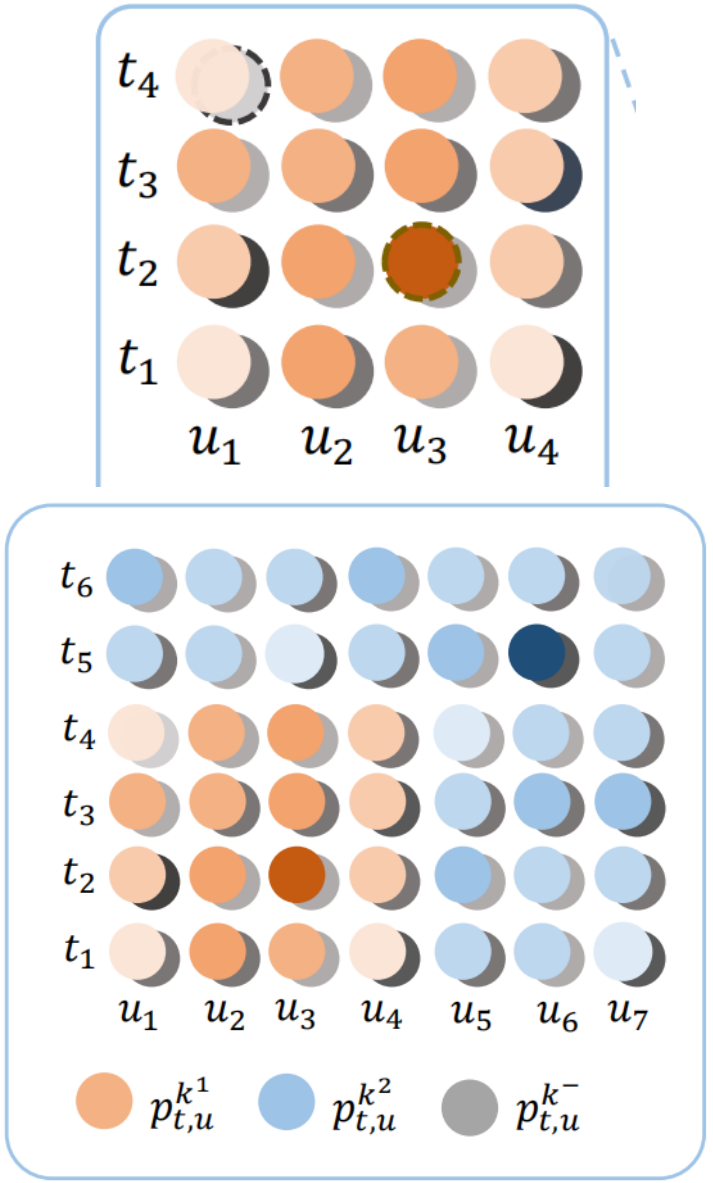
- 现有方法存在问题
  - 现有方法对非情绪帧和情绪帧区分不清；
  - 现有大多数方法也难以充分利用文本信息
- 解决方法
  - 最大池化情绪损失

$$L_{lattice} = -log \max_{t,u} p_{t,u}^{k^*} - log \min_{t,u} p_{t,u}^{k^-}$$

其中 $p_{t,u}^{k^*}$表示当前情绪最强的位置， $p_{t,u}^{k^-}$表示具有最大非情绪或中性类别概率的节点

对于每一帧t，在所有u上做最大池化

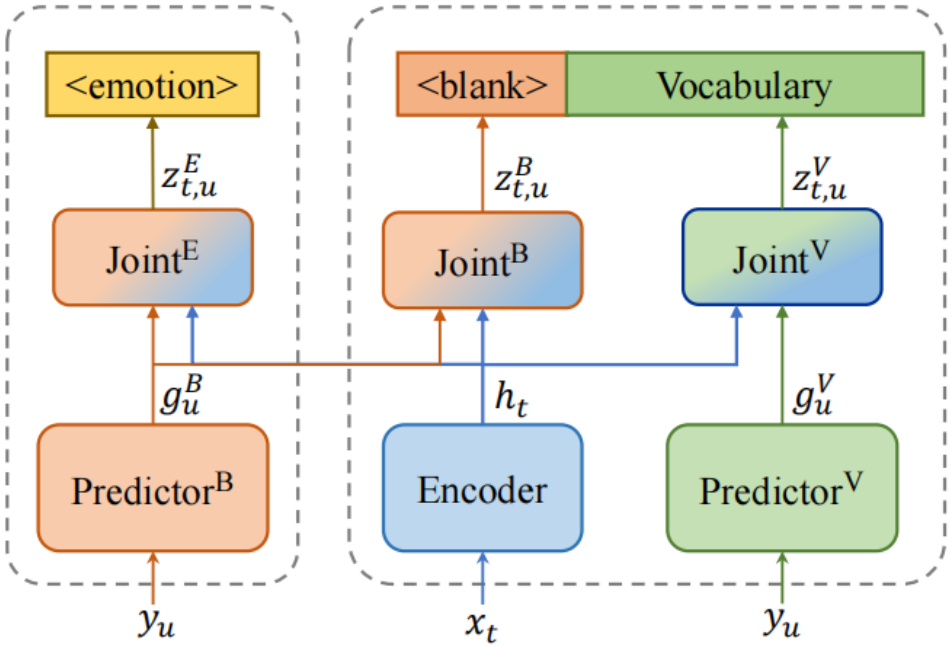
$$y_t = arg \max_{t} \max_{u} p_{t,u}^{k^*}$$

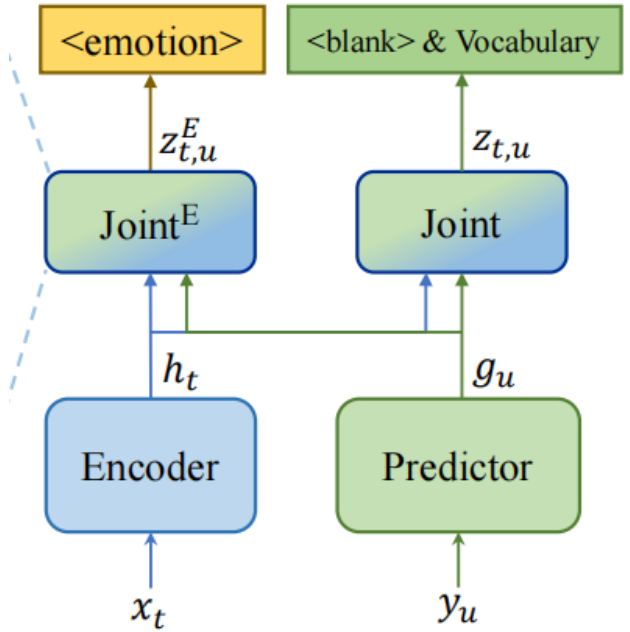


ENT/FENT 创新点分析

- 解决方法
  - 情绪边界建模
    - $Joint^B$:当前不输出 token的同时判断当前是否发生情绪变化
    - $Joint^V$:输出实际转录 token

$$P(y_{\{u+1\}} \mid x_{\{\le t\}}, y_{\{\le u\}}) = softmax\left(\left[z^{\{(B)\}}_{\{t,u\},}; , z^{\{(V)\}}_{\{t,u\}}\right]\right)$$





实验设计　数据资源

- 数据集
  - IEMOCAP（10位演员，5男5女，151个对话，6类情绪，总时长11:28:12）
  - ZED（73个说话者，5位专业标注员标注，180个语句，4类情绪，总时长17分钟）
- 对比方法

  Wav2vec2-PT(2021)、Corr Attentive (2023)、DCW+TsPA (2023)、 Shiftformer (2023)、MSTR (2023)、EmotionNAS (2023)
- 评价指标
  - UA：非加权平均召回率
  - WA：加权平均召回率

实验设计　对比实验

- 评估ENT和FENT在数据集上的表现
  - 在IEMOCAP数据集上ENT 整体性能略高于 FENT
  - ENT在所有方法中效果最好
  - FENT在单句情绪检测中效果在中上

Type	Method	WA (%)	WER (%)
CTC	e2e-ASR [18]	68.60	35.70
	wav2vec 2.0+co-attention [19]	63.40	32.70
RNN-T	Emotion tag [22]	58.20	26.70
ENT	ENT (ours)	**72.43**	26.47
	FENT (ours)	71.84	**25.99**

Method	Year	WA (%)	UA (%)
Wav2vec2-PT [1]	2021	67.90	-
Corr Attentive [2]	2023	-	70.01
DCW+TsPA [3]	2023	72.08	72.17
Shiftformer [4]	2023	72.10	72.70
MSTR [5]	2023	70.60	71.60
EmotionNAS [6]	2023	69.10	72.10
ENT (ours)	2023	**72.43**	**73.88**
FENT (ours)	2023	71.84	72.37

- 评估ENT和FENT的不同模块在数据集上的表现
  - Frame-wise：基于帧的传统情绪识别方法，不使用 Transducer、不建 lattice
  - -w/o $L_{lattice}$：移除最大池化情绪损失（只训练句级 CrossEntropy）
  - -w. $L^{T}_{lattice}$：仅在时间维度 t 上做最大池化监督（横向）
  - -w. $L^{U}_{lattice}$:仅在 token 维度 u 上最大池化监督（纵向）

Method	IEMOCAP		ZED	
	UA ↑	WER ↓	EDER ↓	WER ↓
Frame-wise	68.43	-	59.73	-
ENT	**73.88**	26.47	56.60	39.37
-w/o $\mathcal{L}_{lattice}$	71.76	26.06	62.47	39.19
-w. $\mathcal{L}^{T}_{lattice}$	73.11	26.42	61.88	39.39
-w. $\mathcal{L}^{U}_{lattice}$	69.86	26.14	61.40	**38.82**
-w. $\mathcal{L}^{all}_{lattice}$	71.85	26.19	61.12	39.28
-w. mixing	-	-	52.76*	42.54*
-w. BPE	71.37	30.13	67.68	47.42
FENT	72.37	**25.99**	**55.07**	39.34
-w/o $\mathcal{L}_{lattice}$	71.84	26.69	60.86	39.14
-w. $\mathcal{L}^{T}_{lattice}$	72.52	26.28	59.18	39.48
-w. $\mathcal{L}^{U}_{lattice}$	69.67	26.23	60.86	39.17
-w. $\mathcal{L}^{all}_{lattice}$	69.61	26.18	59.38	39.11
-w. mixing	-	-	54.41*	39.93*
-w. BPE	70.33	30.96	65.63	47.26

实验设计　消融实验

- 评估ENT和FENT的不同模块在数据集上的表现
  - -w. $L_{lattice}^{all}$:在所有 t, u 上平均loss（不做最大池化）
  - -w. mixing:移除 blank 情绪预测策略（不再绑定 blank）
  - -w. BPE:不使用字节对编码（BPE），直接用字符级 token（更细但 WER 高）

Method	IEMOCAP		ZED	
	UA ↑	WER ↓	EDER ↓	WER ↓
Frame-wise	68.43	-	59.73	-
ENT	**73.88**	26.47	56.60	39.37
-w/o $\mathcal{L}_{lattice}$	71.76	26.06	62.47	39.19
-w. $\mathcal{L}_{lattice}^{T}$	73.11	26.42	61.88	39.39
-w. $\mathcal{L}_{lattice}^{U}$	69.86	26.14	61.40	**38.82**
-w. $\mathcal{L}_{lattice}^{all}$	71.85	26.19	61.12	39.28
-w. mixing	-	-	52.76*	42.54*
-w. BPE	71.37	30.13	67.68	47.42
FENT	72.37	**25.99**	**55.07**	39.34
-w/o $\mathcal{L}_{lattice}$	71.84	26.69	60.86	39.14
-w. $\mathcal{L}_{lattice}^{T}$	72.52	26.28	59.18	39.48
-w. $\mathcal{L}_{lattice}^{U}$	69.67	26.23	60.86	39.17
-w. $\mathcal{L}_{lattice}^{all}$	69.61	26.18	59.38	39.11
-w. mixing	-	-	54.41*	39.93*
-w. BPE	70.33	30.96	65.63	47.26

算法总结 ENT/FENT

- 算法贡献
  - 构建情感联合网络：模型能在任意点输出情绪分布，实现帧级/词级的细粒度情绪建模
  - 设计格最大池化损失：在没有帧级标签的情况下，仍能识别出“哪一帧情绪最强”
  - 情绪边界建模：利用 ASR 中的 <blank> 符号，帮助模型识别情绪变化的边界位置

- 算法不足
  - 模块较多，训练成本高，调参繁琐
  - 跨句上下文建模能力弱
  - ZED 数据规模小，泛化能力尚未充分验证

特点总结与未来展望

特点总结与未来展望

- 特点总结
  - PCGNet
    - 构建人格感知图结构，同时构建情绪识别和情绪转变多任务图结构，增加情绪转换对比学习
      - 考虑说话者与听者之间的互动存在差异，情绪变化和情绪变化类型
  - ENT/FENT
    - 构建情感联合网络，设计格最大池化损失，情绪边界建模
      - 有效进行了情绪变化的识别，实现帧级/词级的细粒度情绪建模
- 未来发展
  - 如何将帧级精细感知能力 + 上下文/人格建模能力相结合
  - 如何处理多种语言泛化的问题
  - 如何更好的结合多模态，充分利用其特征提升准确率的问题

参考文献

[1] Tu, Geng, et al. "A Persona-Infused Cross-Task Graph Network for Multimodal Emotion Recognition with Emotion Shift Detection in Conversations." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

[2] Shen, Siyuan, et al. "Emotion neural transducer for fine-grained speech emotion recognition." ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024.

附录

- 数据集介绍

数据集名称	类型	场景	平均句时长	平均词数	情绪标签种类	标签粒度	总时长	utterance数	对话数
IEMOCAP	多模态（音频+文本+视频）	对话（双人）	≈ 4.5 秒	≈ 11	10类（常用6类）	utterance级	≈ 12 小时	≈ 10,000	≈ 150
MELD	多模态（音频+文本+视频）	多说话者对话	≈ 3.5 秒	≈ 8	7类	utterance级	≈ 13 小时	≈ 13,000	≈ 1,400
EmoryNLP	文本	多说话者对话	—	≈ 15	7类（细粒度）	utterance级	—	≈ 14,000	≈ 900
DailyDialog	文本	对话（双人）	—	≈ 10	7类	utterance级	—	≈ 100,000	≈ 1,3000
ZED	音频	自然语音+情绪段	≈ 5~10 秒	≈ 12~25	4类（带时间段）	时间段级	17分钟	≈ 180	—

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！