# 成都信息工程大学

## 本科毕业设计开题报告

**题目：基于多模态情感分析的研究和应用**

学　　院：　计算机学院

专　　业：　计算机科学与技术

学生姓名：　梁嘉轩

指导教师：　冯翱

日　　期：　2026 年 1 月 5 日

# 目录

# 1 课题背景与意义

## 1.1 研究背景

随着移动互联网的高速发展，**短视频平台用户规模已突破 10 亿**，以抖音、快手、B 站为代表的内容平台每日产生海量多模态数据。传统的**单模态情感分析**方法在面对复杂语义时暴露出显著局限性：

- **文本模态局限**：纯文本难以准确识别**反讽**（如" 真是太棒了" 的负面含义）与**双关**等修辞手法，缺乏语气、表情等辅助信息。
- **视觉模态局限**：单纯的面部表情识别在非受控环境下易受光照、遮挡影响，且无法区分" 职业微笑" 与真实情感。
- **声学模态局限**：孤立的语音信号难以判断说话者的真实意图，尤其在背景噪声干扰时鲁棒性较差。

一个典型案例是中文语境下的**" 笑哭" 表情（ ）**：用户常以此表达无奈、嘲讽甚至悲伤，与字面" 笑" 的含义形成矛盾。单一信息源无法解析这种**跨模态语义冲突**，导致情感判断失误。

## 1.2 研究意义

本课题通过融合**文本、视觉、声学三通道**信息，旨在解决上述单模态瓶颈，具有以下核心价值：

1. **互补性验证**：当某一模态信息模糊时（如文字反讽），其他模态（如语气愤怒、表情严肃）可提供印证，消除歧义。
2. **鲁棒性提升**：多通道冗余设计使模型在部分信息缺失（如视频无字幕、音频静音）时仍能维持基本判断能力。
3. **应用场景广泛**：研究成果可直接应用于**舆情监控**（识别网络负面情绪）、**智能客服**（判断用户满意度）及**人机交互**（情感陪伴机器人）等领域。

# 2 国内外研究现状

## 2.1 技术发展历程

多模态情感分析的研究经历了三个主要阶段：

1. **早期拼接阶段 (2015 前)**：采用 **Early Fusion**（特征级拼接）或 **Late Fusion**（决策级融合）的简单策略，将各模态特征直接串联后送入分类器。该方法实现简单，但忽略了模态间的交互关系。
2. **中期注意力阶段 (2015-2019)**：引入**双流网络**与**注意力机制**，允许模型动态学习模态权重。代表性工作如 TFN (Tensor Fusion Network) 和 MFN (Memory Fusion Network) 尝试捕捉模态间的高阶交互。
3. **当前大模型阶段 (2020 至今)**：基于 **Transformer** 架构的预训练模型成为主流。**CLIP** (OpenAI) 实现了视觉-语言的跨模态对齐；**ViT** (Vision Transformer) 将图像建模为序列；**BERT** 及其变体（如中文 RoBERTa）在文本理解上取得突破。这些预训练模型为下游多模态任务提供了强大的特征提取能力。

## 2.2 当前主要挑战

尽管技术快速发展，多模态情感分析仍面临以下核心难题：

- **模态对齐困难 (Alignment)**：文本、视频、音频的时序粒度不同（词级 vs 帧级 vs 音素级），如何实现精准的跨模态时序对齐仍是开放问题。

- **异构特征融合**：不同模态的特征空间差异巨大（如 BERT 输出 768 维语义向量，ResNet 输出 2048 维视觉特征），简单拼接可能导致维度爆炸或信息稀释。
- **中文数据集稀缺**：现有主流数据集（如 CMU-MOSI, CMU-MOSEI）以英文为主，**高质量中文多模态情感数据集极度匮乏**，这直接制约了中文场景下的模型训练与评估。

上述挑战凸显了选择合适中文数据集的重要性，这也是本项目选用 **CH-SIMS** 数据集的核心动因。

# 3 数据集介绍

## 3.1 CH-SIMS 数据集概述

本项目选用 **CH-SIMS** (Chinese Single- and Multi-modal Sentiment Analysis) 数据集，这是**首个包含细粒度标注的中文多模态情感数据集**，由清华大学于 ACL 2020 发布。

- **数据来源**：非受控环境下的**影视片段**（电影、电视剧、综艺），涵盖多样化的说话人、场景和情感表达。
- **样本规模**：共 **2,281** 个标注样本，划分为训练集 (1,368) / 验证集 (456) / 测试集 (457)。
- **标注特点**：提供**多模态独立标注**，即文本、视觉、声学三个模态分别有独立的情感标签，便于分析各模态的贡献度。

## 3.2 数据集难点

- **类别非平衡**：正负情感样本分布不均，需引入 **Class Weights**（类别加权）或过采样策略以防止模型偏向多数类。
- **噪声干扰**：部分样本存在背景音乐、多人对话、画面遮挡等干扰因素，对模型鲁棒性提出较高要求。

## 3.3 数据预处理流程

为确保多模态特征的有效提取，本项目设计以下预处理管线：

1. **时序对齐**：采用 CH-SIMS 官方提供的 **Word-level Alignment** 标注，将视频帧与音频片段对齐至文本词级粒度。
2. **人脸检测与对齐**：
   - 使用 **MTCNN** (Multi-task Cascaded Convolutional Networks) 检测人脸边界框及 **5 点关键点**（双眼、鼻尖、嘴角）。
   - 基于关键点进行**仿射变换**，将人脸归一化至统一尺寸 ($224 \times 224$)，消除姿态与尺度差异。
3. **数据清洗**：
   - 剔除**黑帧**（全黑或无有效内容的视频帧）。
   - 移除**静音片段**（音频能量低于阈值的样本）。
   - 过滤人脸检测失败的样本。

# 4 核心技术路线

## 4.1 总体架构

本研究采用**三塔架构 (Three-Tower Architecture)** 进行多模态特征提取，并通过 **Early Fusion（特征级拼接）**实现模态融合。该方案的核心优势在于：结构简洁、易于调试、显存占用可控，适合在有限硬件条件下完成毕业设计。

整体流程为：**三路独立编码** → **特征拼接** → **全连接分类**。

## 4.2 文本塔 (Text Tower)

- **模型选择**：采用 `bert-base-chinese` 预训练模型（HuggingFace 官方发布）。
- **特征提取**：将输入文本经 Tokenizer 编码后送入 BERT，取最后一层的 **[CLS] token** 作为句级语义表示。
- **输出维度**：**768** 维向量。

## 4.3 视觉塔 (Visual Tower)

- **人脸预处理**：使用 **MTCNN** 进行人脸检测，提取 **5 点关键点**（双眼中心、鼻尖、左右嘴角），通过仿射变换完成人脸对齐，输出 $224 \times 224$ 归一化图像。
- **特征提取**：将对齐后的人脸帧序列送入预训练的 **ResNet-50**（在 ImageNet 上预训练），提取每帧的 2048 维特征向量。
- **时序聚合**：对视频片段内的所有帧特征进行**时序均值池化 (Temporal Mean Pooling)**：

$$\mathbf{V} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{v}_i \tag{1}$$

  其中 $N$ 为帧数，$\mathbf{v}_i$ 为第 $i$ 帧的 ResNet 输出。

- **输出维度**：**2048** 维向量。

## 4.4 声学塔 (Acoustic Tower)

- **特征提取**：使用 **OpenSMILE** 工具包提取底层声学特征，包括：
  - **MFCC**（梅尔频率倒谱系数）：13 维 + $\Delta$ + $\Delta\Delta$ = 39 维
  - **Chroma**（色度特征）：12 维
  - **Energy**（能量）、**Pitch**（基频）等韵律特征
- **时序编码**：将帧级声学特征送入 **Conv1D** 卷积层进行局部时序建模。
- **聚合策略**：经卷积后进行 **MeanPool**，输出固定长度向量。
- **输出维度**：**256** 维向量。

## 4.5 融合与分类

- **特征拼接**：将三塔输出直接拼接 (Concatenation)：

$$\mathbf{F} = [\mathbf{T}; \mathbf{V}; \mathbf{A}] \in R^{768+2048+256} = R^{3072} \tag{2}$$

- **正则化**：引入 **Dropout**($p = 0.3$) 防止过拟合。
- **分类层**：拼接特征经全连接层 (FC) 映射至类别数，输出层使用 **Softmax** 激活。
- **损失函数**：**CrossEntropyLoss**，结合 Class Weights 处理类别不平衡。

### 4.6 硬件可行性与显存优化

本项目的硬件环境为 **RTX 3060 Laptop GPU (6GB 显存)**，为确保模型可在该配置下顺利训练，采用以下优化策略：

1. **梯度累积 (Gradient Accumulation)**：将大 Batch 拆分为多个小 Batch，累积梯度后统一更新参数，等效增大批量而不增加显存峰值。
2. **冻结骨干网络 (Freeze Backbone)**：冻结 BERT 和 ResNet 的底层参数，仅微调顶层，大幅减少可训练参数量与显存占用。
3. **混合精度训练 (FP16)**：使用 PyTorch AMP 自动混合精度，在保持精度的前提下降低显存消耗约 30%–50%。

## 5 实验设计与预期产出

### 5.1 消融实验设计

为验证各模态的独立贡献及多模态融合的有效性，本项目设计以下**消融实验 (Ablation Study)**：

表 1: 消融实验配置

| 实验组 | 模态组合 | 验证目标 |
|---|---|---|
| Baseline-T | 仅文本 (Text) | 文本模态独立性能 |
| Baseline-V | 仅视觉 (Visual) | 视觉模态独立性能 |
| Baseline-A | 仅声学 (Acoustic) | 声学模态独立性能 |
| Fusion-TV | 文本 + 视觉 | T-V 互补性验证 |
| Fusion-TA | 文本 + 声学 | T-A 互补性验证 |
| Fusion-VA | 视觉 + 声学 | V-A 互补性验证 |
| Full Model | 文本 + 视觉 + 声学 | 三模态融合最优性能 |

### 5.2 评估指标

- **Accuracy (准确率)**：正确分类样本占总样本的比例，衡量整体分类性能。
- **Weighted F1 (加权 F1 值)**：考虑类别不平衡的综合指标，对各类别的 F1 按样本数加权平均，更适合非平衡数据集评估。

### 5.3 预期实验结论

1. 单模态中，**文本模态 (Baseline-T)** 预期表现最优，因语义信息最为直接。
2. 双模态融合相比单模态有显著提升，验证模态间的**互补性假设**。
3. **Full Model**（**三模态融合**）预期达到最优性能，证明三通道信息的综合价值。

### 5.4 预期产出

1. **工程代码**：基于 PyTorch 的完整多模态情感分析系统源码，包含数据预处理、模型定义、训练与评估脚本。
2. **可视化演示**：提供简易 Demo 界面，支持输入视频片段并输出情感预测结果。
3. **学术论文**：包含消融实验分析、性能对比及结论讨论的本科毕业设计论文。

## 6 进度安排

表 2: 毕业设计进度计划

| 时间 | 主要任务 |
|---|---|
| **1–2 月** | • CH-SIMS 数据集下载与格式解析<br>• MTCNN 人脸检测与关键点对齐<br>• 数据清洗（黑帧剔除、静音过滤）<br>• OpenSMILE 声学特征提取 |
| **3 月** | • 搭建三塔架构模型（BERT / ResNet / Conv1D）<br>• 实现 Early Fusion 拼接与分类层<br>• 初步训练与 Debug，验证数据流通畅<br>• 显存优化调试（梯度累积、Freeze Backbone） |
| **4 月** | • 完成 7 组消融实验（单模态 / 双模态 / 全模态）<br>• 超参数调优（学习率、Dropout、Batch Size）<br>• 整理实验数据，绘制对比图表 |
| **5 月** | • 撰写毕业设计论文<br>• 准备答辩 PPT 与演示 Demo<br>• 代码整理与文档归档 |

## 参考文献

[1] Yu, W., et al. "CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset." ACL 2020.

[2] Devlin, J., et al. "BERT: Pre-training of Deep Bidirectional Transformers." NAACL 2019.