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

随着移动互联网与 5G 技术的普及，以抖音、快手、B 站为代表的**短视频平台用户规模已突破 10 亿**，人类的信息表达方式已从单一文本彻底转向包含**文本、图像、语音**的多模态形式。每日数以亿计的短视频、直播弹幕与语音评论构成了巨量的非结构化情感数据，对其进行准确的情感分析具有重要的学术与商业价值。

然而，传统的情感分析技术主要依赖**单一模态**（如仅分析评论文字），在面对复杂的网络表达时存在显著局限性：

- **反讽识别失效**：纯文本模型难以理解” 真是太棒了” 在特定语境下表达的是讽刺而非赞美，因为反讽的判断往往依赖语气、表情等非文本线索。
- **双关语义模糊**：中文的谐音梗、网络流行语（如” 绝绝子”” 无语子”）含有多义性，单纯的词向量模型难以消歧。
- **表情包矛盾**：在中文语境下，用户常使用**” 笑哭” 表情**（ ）配合负面文字，或用” 微笑” 表情表达不满。此时，表情的字面含义与真实情感完全相悖，单一视觉或文本模型极易产生误判。

此外，现有的多模态研究多基于英文数据集（如 CMU-MOSI、CMU-MOSEI），**高质量的中文多模态情感数据集相对稀缺**。中文在语义结构（缺乏形态变化、依赖语序）、语音韵律（四声调）上与英文存在巨大差异，直接迁移国外模型往往效果不佳。这一现状迫切要求我们构建适配中文语境的多模态情感分析方案。

1.2 研究意义

本课题旨在构建一个适配中文语境的多模态情感分析系统，具有重要的**理论价值**与**应用前景**：

1.2.1 理论价值

1. **模态互补性验证**：通过融合文本、视觉、声学三通道信息，验证多模态数据的互补效应。当文字语义模糊时（如反讽），面部表情（皱眉、撇嘴）和语音语调（尖锐、低沉）可提供辅助判断依据，实现**跨模态消歧**。
2. **鲁棒性提升**：多通道设计具备**冗余容错**能力。当某一模态信息缺失（如视频无字幕、音频静音、画面遮挡）时，其他模态仍可维持基本的情感判断，避免系统完全失效。
3. **中文场景适配**：针对中文的语义特点与韵律特征进行专项优化，填补国内在中文多模态情感分析领域的研究空白。

1.2.2 应用前景

1. **舆情监控**：实时分析社交媒体上的短视频与评论情感，识别潜在的负面舆论热点，为政府与企业提供预警。
2. **智能客服**：通过分析用户的语音语调与面部表情，判断客户满意度与情绪状态，辅助客服人员进行针对性回应。
3. **人机交互**：为情感陪伴机器人、智能家居助手等提供” 情商” 能力，使机器能够感知用户的情绪变化并做出恰当反馈。

综上所述，本课题不仅具有填补中文多模态情感分析研究空白的学术意义，更具备广阔的产业落地前景，为构建高” 情商” 的智能系统提供关键技术支撑。

2 国内外研究现状

2.1 技术发展历程

多模态情感分析作为自然语言处理与计算机视觉的交叉领域，经历了三个主要发展阶段：

2.1.1 早期融合阶段 (2010–2015)

早期研究采用**简单特征拼接**策略，将各模态的手工特征（如文本的 TF-IDF、视觉的 HOG、声学的 MFCC）直接串联，送入 SVM 或逻辑回归等浅层分类器。代表性方法包括 **Early Fusion**（特征级融合）与 **Late Fusion**（决策级融合）。这类方法实现简单，但忽略了模态间的交互关系，且对手工特征的设计高度依赖。

2.1.2 张量融合阶段 (2016–2019)

随着深度学习的发展，研究者开始探索更复杂的融合机制。Zadeh 等人提出的 **TFN (Tensor Fusion Network)** 通过外积运算捕捉模态间的高阶交互；**MFN (Memory Fusion Network)** 引入记忆机制建模时序依赖。这一阶段的方法显著提升了融合效果，但计算复杂度较高，且张量膨胀问题限制了其扩展性。

2.1.3 预训练大模型阶段 (2020 至今)

基于 **Transformer** 架构的预训练模型成为当前主流范式：

- **BERT** [2]：通过双向语言模型预训练，在文本理解任务上取得突破性进展，其中文版本 `bert-base-chinese` 成为中文 NLP 的基础设施。
- **ViT (Vision Transformer)**：将图像建模为 Patch 序列，打破了 CNN 在视觉领域的垄断。
- **CLIP**：OpenAI 提出的视觉-语言对比学习模型，实现了图文的跨模态对齐。

这些预训练模型为下游多模态任务提供了强大的**特征提取器**，但其庞大的参数量也带来了显存与算力挑战。

2.2 国内外研究现状对比

2.2.1 国外研究

国外研究以英文数据集为主，形成了较为完善的基准体系：

- **CMU-MOSI**：卡内基梅隆大学发布的英文多模态情感数据集，包含 2,199 个视频片段，是该领域最经典的基准之一。
- **CMU-MOSEI**：MOSI 的扩展版，样本量达 23,000+，覆盖更丰富的情感标签。
- **IEMOCAP**：多模态情感对话数据集，常用于对话情感识别研究。

2.2.2 国内研究

相比之下，**中文多模态情感数据集极度稀缺**，主要痛点包括：

- 现有中文情感数据集多为单模态（如微博文本、电商评论），缺乏对应的视频与音频标注。
- 中文的语义表达方式（成语、网络流行语、谐音梗）与英文差异显著，直接迁移英文预训练模型效果有限。
- 中文语音的四声调韵律信息对情感判断有重要影响，但现有声学特征提取工具多针对英语优化。

2.3 当前核心挑战

综合国内外研究现状，多模态情感分析仍面临以下**核心挑战**：

1. **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），精确的跨模态时序对齐仍是开放问题。

2. **异构特征融合**：不同模态的特征空间差异巨大（BERT 输出 768 维语义向量 vs ResNet 输出 2048 维视觉特征），简单拼接可能导致信息冗余或特征互相干扰。

3. **训练资源消耗**：BERT + ResNet 等大模型的联合训练对显存要求极高，限制了其在普通硬件环境下的应用。

上述挑战凸显了选择合适中文数据集与设计轻量化融合方案的重要性，这也是本项目选用 **CH-SIMS** 数据集并采用**三塔架构 + Early Fusion** 策略的核心动因。

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) 数据集作为实验基准。该数据集由清华大学于 ACL 2020 发布 [1]，是**首个包含细粒度独立标注的中文多模态情感数据集**，填补了中文多模态情感分析领域的数据空白。

3.1.1 数据来源

CH-SIMS 的样本来源于**非受控环境**下的影视片段，包括电影、电视剧、综艺节目等。与实验室采集的受控数据不同，这些真实场景样本具有以下特点：

- 说话人多样：涵盖不同年龄、性别、口音的演员。
- 场景复杂：包含多人对话、背景音乐、光照变化等干扰因素。
- 情感表达自然：非刻意表演，接近日常生活中的情感流露。

3.1.2 数据规模与划分

表 1: CH-SIMS 数据集划分

集合	训练集	验证集	测试集
样本数	1,368	456	457
占比	60%	20%	20%

总样本量为 **2,281** 个视频片段，每个片段时长 2–8 秒，配有对应的文字转录、视频帧与音频轨道。

3.1.3 标注特点

CH-SIMS 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（正向/中性/负向）。这一设计使研究者能够：

- 分析各模态对最终情感判断的**独立贡献**。
- 研究模态间的**一致性与冲突**现象（如文字积极但语气消极）。

3.2 数据集难点

1. **类别非平衡**：正、中、负三类样本分布不均，需引入 **Class Weights**（类别加权损失）防止模型偏向多数类。

2. **噪声干扰**：部分样本存在背景音乐、多人重叠说话、人脸遮挡等问题，对模型鲁棒性提出较高要求。

3. **模态缺失**：少量样本存在单模态信息缺失（如无字幕、静音片段），需在预处理阶段进行筛选或填充。

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**三阶段预处理管线**：

3.3.1 时序对齐

采用 CH-SIMS 官方提供的 **Word-level Alignment** 标注文件，将视频帧序列与音频片段精确对齐至文本的词级粒度。对齐信息记录了每个词对应的起止时间戳，便于后续按词切分视觉与声学特征。

3.3.2 视觉预处理

1. **人脸检测**：使用 **MTCNN** (Multi-task Cascaded Convolutional Networks) 检测视频帧中的人脸边界框。
2. **关键点定位**：提取人脸的 **5 点关键点**（左眼中心、右眼中心、鼻尖、左嘴角、右嘴角），用于后续对齐。
3. **仿射变换**：基于关键点计算仿射变换矩阵，将人脸归一化至 224×224 的标准尺寸，消除姿态、尺度差异。
4. **黑帧剔除**：检测并移除全黑或无有效内容的帧（通过计算帧像素方差，低于阈值则判定为无效帧）。

3.3.3 声学预处理

1. **特征提取**：使用 **OpenSMILE** [4] 工具包提取标准化的底层声学特征，包括：
   - **MFCC**（梅尔频率倒谱系数）：13 维静态系数 + 一阶差分 + 二阶差分 = 39 维
   - **Chroma**（色度特征）：12 维，反映音高分布
   - **Energy**（能量）、**Pitch**（基频）等韵律特征
2. **静音过滤**：检测音频能量低于阈值的静音片段，并进行标记或剔除。
3. **特征归一化**：对提取的声学特征进行 Z-Score 标准化，消除量纲差异。

4 核心技术路线

4.1 总体架构设计

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。该方案遵循” 分而治之” 的设计哲学：

设计理念：三路独立编码 → 特征拼接 → 联合分类

相比复杂的张量融合或跨模态注意力机制，Early Fusion 具有以下**工程优势**：

- **结构简洁**：各模态编码器独立，易于调试与维护。
- **显存可控**：无需存储高阶交互张量，适合在消费级 GPU 上运行。
- **可解释性**：通过消融实验可清晰量化各模态的贡献。

4.2 文本塔 (Text Tower)

4.2.1 模型选择

采用 `bert-base-chinese` 预训练模型（由 Google 发布，HuggingFace 托管），该模型在大规模中文语料上进行了 Masked Language Model 预训练，具备强大的中文语义理解能力。

4.2.2 特征提取流程

1. 将输入文本经 **BertTokenizer** 分词，添加 [CLS] 和 [SEP] 特殊标记。
2. 将 Token 序列送入 BERT 编码器，经过 12 层 Transformer Block。
3. 取最后一层的 **[CLS] token 隐藏状态**作为句级语义表示。

4.2.3 输出规格

$$\mathbf{T} \in R^{768}$$

其中 768 为 BERT-base 的隐藏层维度。

4.3 视觉塔 (Visual Tower)

4.3.1 人脸预处理

使用 **MTCNN** 进行人脸检测与关键点定位：

- 检测人脸边界框 (Bounding Box)。
- 提取 **5 点面部关键点**：双眼中心、鼻尖、左右嘴角。
- 基于关键点进行**仿射变换**，将人脸归一化至 224×224。

4.3.2 特征提取

将对齐后的人脸帧序列送入预训练的 **ResNet-50** [3]（在 ImageNet 上预训练），提取每帧的高层视觉特征：

- 移除 ResNet 的最后全连接层，取 **AvgPool 输出**作为帧级特征。
- 每帧输出 **2048 维**特征向量。

4.3.3 时序聚合

对视频片段内的 N 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 $\mathbf{v}_i \in R^{2048}$ 为第 i 帧的 ResNet 输出。

设计说明：采用均值池化而非循环网络（LSTM/GRU）的原因是：

- **显存节约**：均值池化无可训练参数，显著降低显存占用。
- **计算高效**：避免序列依赖的反向传播，加速训练。
- **鲁棒性**：对帧数变化不敏感，自动适应不同长度的视频。

4.3.4 输出规格

$$\mathbf{V} \in R^{2048}$$

4.4 声学塔 (Acoustic Tower)

4.4.1 特征提取

使用 **OpenSMILE** [4] 提取帧级声学特征，主要包括：

- **MFCC**：39 维（13 维静态 + Δ + ΔΔ）
- **Chroma**：12 维
- **Energy / Pitch**：韵律特征

4.4.2 时序编码

将帧级声学特征序列送入 **一维卷积网络 (Conv1D)**：

- **Conv1D**：提取局部时序模式，捕捉语音中的短时情感波动。
- **ReLU 激活**：引入非线性。
- **MeanPool**：对时序维度进行均值池化，输出固定长度向量。

4.4.3 输出规格

$$\mathbf{A} \in R^{256}$$

4.5 模态融合与分类

4.5.1 Early Fusion 拼接

将三塔输出在特征维度上**直接拼接 (Concatenation)**：

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

4.5.2 分类头

1. **Dropout**：$p = 0.3$，随机丢弃神经元以防止过拟合。
2. **全连接层**：$\mathrm{FC} : R^{3072} \to R^{C}$，其中 C 为类别数。
3. **Softmax**：输出各类别的概率分布。

4.5.3 损失函数

采用**加权交叉熵损失 (Weighted Cross-Entropy)**：

$$\mathcal{L} = -\sum_{c=1}^{C} w_c \cdot y_c \log(\hat{y}_c) \tag{3}$$

其中 w_c 为类别权重，用于缓解数据不平衡问题。

4.6 硬件适配与显存优化

本项目的开发环境为 **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**轻量化策略**：

4.6.1 梯度累积 (Gradient Accumulation)

将逻辑 Batch Size 拆分为多个小批次：

- 设物理 Batch Size = 4，累积步数 = 8。
- 等效 Batch Size = $4 \times 8 = 32$。
- 每 8 步执行一次参数更新，显存峰值仅需容纳 4 个样本。

4.6.2 冻结骨干网络 (Freeze Backbone)

- **第一阶段**：冻结 BERT 和 ResNet 的全部参数，仅训练分类头与声学塔。
- **第二阶段**：解冻顶层（如 BERT 最后 2 层），进行端到端微调。

该策略可将可训练参数量从约 1.5 亿降至约 500 万，显存占用降低 70% 以上。

4.6.3 混合精度训练 (FP16)

使用 PyTorch AMP (Automatic Mixed Precision)：

- 前向传播使用 FP16，反向传播自动缩放梯度。
- 在保持训练精度的前提下，显存消耗降低约 30%–50%。

5 实验设计与预期产出

5.1 消融实验设计

为系统验证各模态的**独立贡献**及多模态融合的**互补效应**，本项目设计以下 **7 组消融实验 (Ablation Study)**：

表 2: 消融实验配置

序号	实验组	模态组合	验证目标
1	Baseline-T	仅文本 (Text)	文本模态独立性能上界
2	Baseline-V	仅视觉 (Visual)	面部表情对情感判断的贡献
3	Baseline-A	仅声学 (Acoustic)	语音韵律对情感判断的贡献
4	Fusion-TV	文本 + 视觉	T-V 互补性验证
5	Fusion-TA	文本 + 声学	T-A 互补性验证
6	Fusion-VA	视觉 + 声学	V-A 互补性验证（无文本基线）
7	Full Model	文本 + 视觉 + 声学	三模态融合最优性能

5.2 实验假设

基于多模态互补性理论，本项目提出以下**预期假设**：

1. **H1（文本主导假设）**：在单模态实验中，Baseline-T 将取得最高性能，因为文本语义信息最为直接。
2. **H2（模态互补假设）**：任意双模态组合的性能将**优于**各自的单模态基线，证明模态间存在互补效应。
3. **H3（三模态最优假设）**：Full Model 将达到实验中的**最优性能**，验证三通道信息融合的综合价值。

5.3 评估指标

- **Accuracy (准确率)：**

$$\text{Accuracy} = \frac{\text{正确预测数}}{\text{总样本数}} \tag{4}$$

衡量模型的整体分类正确率。

- **Weighted F1 (加权 F1 值)：**

$$\text{Weighted F1} = \sum_{c=1}^{C} \frac{n_c}{N} \cdot F1_c \tag{5}$$

其中 n_c 为类别 c 的样本数，N 为总样本数。该指标对各类别的 F1 按样本数加权平均，**更适合评估非平衡数据集**。

5.4 实验环境

表 3: 实验环境配置

组件	规格
GPU	NVIDIA RTX 3060 Laptop (6GB)
CPU	Intel Core i7-12700H
内存	16GB DDR5
操作系统	Windows 11 / Ubuntu 22.04 (WSL2)
深度学习框架	PyTorch 2.0+
预训练模型	bert-base-chinese, ResNet-50 (ImageNet)

5.5 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、训练脚本、评估脚本。
   - 代码结构清晰，配有详细注释与 README 文档。
2. **可视化演示：**
   - 提供基于 Gradio 或 Streamlit 的简易 Demo 界面。
   - 支持上传短视频片段，实时输出情感预测结果与置信度。
3. **学术论文：**
   - 包含完整消融实验分析的本科毕业设计论文。
   - 论文结构符合学校规范，涵盖背景、方法、实验、结论等章节。

6 进度安排

本毕业设计预计于 2026 年 1 月至 5 月完成，具体进度安排如下：

表 4: 毕业设计进度计划

阶段	时间	主要任务
数据准备	1–2 月	• CH-SIMS 数据集下载与格式解析 • MTCNN 人脸检测与关键点对齐跑通 • 黑帧剔除、静音过滤等数据清洗 • OpenSMILE 声学特征提取与归一化 • 数据加载器 (DataLoader) 封装
模型搭建	3 月	• 搭建三塔架构（BERT / ResNet / Conv1D） • 实现 Early Fusion 拼接与分类头 • 初步训练与 Debug，验证数据流通畅 • 显存优化调试（梯度累积、Freeze Backbone） • 训练日志可视化（TensorBoard / WandB）
实验验证	4 月	• 完成 7 组消融实验（单/双/全模态） • 超参数调优（学习率、Dropout、Batch Size） • 整理实验数据，绘制对比图表 • 分析各模态贡献度，验证互补性假设
论文撰写	5 月	• 撰写毕业设计论文（背景/方法/实验/结论） • 准备答辩 PPT 与可视化 Demo • 代码整理、注释完善与文档归档 • 预答辩演练与修改优化

6.1 里程碑节点

- **2 月底**：完成数据预处理管线，MTCNN + OpenSMILE 流程可复现。
- **3 月底**：三塔模型可在 RTX 3060 上正常训练，验证集 Loss 收敛。
- **4 月中旬**：完成全部 7 组消融实验，整理实验结果表格。
- **5 月中旬**：完成论文初稿，提交导师审阅。
- **5 月底**：完成论文终稿与答辩准备。

参考文献

[1] Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., Jiang, J., and Yang, S. *CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotations of Modality*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 3718–3727.

[2] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019, pp. 4171–4186.

[3] He, K., Zhang, X., Ren, S., and Sun, J. *Deep Residual Learning for Image Recognition*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[4] Eyben, F., Wöllmer, M., and Schuller, B. *openSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor*. Proceedings of the 18th ACM International Conference on Multimedia (MM), 2010, pp. 1459–1462.

[5] Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. *Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks*. IEEE Signal Processing Letters, 2016, 23(10): 1499–1503.

[6] Poria, S., Cambria, E., Bajpai, R., and Hussain, A. *A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion*. Information Fusion, 2017, 37: 98–125.

[7] Zadeh, A., Zellers, R., Pincus, E., and Morency, L. P. *MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos*. IEEE Intelligent Systems, 2016, 31(6): 82–88.

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. *Attention Is All You Need*. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.