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 技术的深度普及，以抖音、快手、B 站为代表的短视频平台用户规模已突破十亿量级，人们的情感表达方式也从单一的文字评论演变为融合表情包、语音弹幕与面部反应的多模态形式。然而，传统的情感分析技术仍以**单模态文本处理**为主导，这在面对中文互联网中高度复杂的表达习惯时显得力不从心。

一个典型的困境是：当用户发送" 你真行" 这句话并配上一个翻白眼的表情包时，纯文本模型几乎必然将其误判为积极情感——因为从字面语义看，" 你真行" 本身确实是一句肯定的表述。然而任何中文母语者都能轻易识别出，这其实是一种典型的**反讽**表达。类似的案例在中文网络空间中俯拾皆是：B 站弹幕中的" 友军厚葬" 看似在赞美队友，实则暗含嘲讽；微信聊天中的" 笑哭" 表情（ ）更是常被用于表达无奈甚至悲伤，与其字面" 大笑" 的含义形成诡异的背离。这些现象揭示了一个核心问题：**当语言本身充满歧义时，仅凭文本信息无法还原说话者的真实意图**。

与此同时，当前高质量的多模态情感数据集主要以英文为主，如 CMU-MOSI 和 CMU-MOSEI，而中文在语义结构、谐音梗使用以及四声调韵律上与英文存在根本性差异。这意味着直接迁移国外预训练模型往往效果欠佳，也凸显了构建适配中文语境的多模态分析方案的紧迫性。

1.2 研究意义

本课题的核心价值在于通过引入**视觉与声学模态**，为文本语义歧义提供额外的消歧依据。具体而言，当文字表达模糊时，说话者的面部表情（皱眉、撇嘴）与语音语调（尖锐、低沉）往往能够暴露其真实情感倾向。这种**模态间的互补性**正是多模态分析相较于单模态方法的本质优势所在。

此外，多通道架构还具备天然的**鲁棒性**：在实际应用场景中，视频可能存在静音片段、无字幕或人脸遮挡等情况，多模态系统可以依赖剩余模态维持基本判断能力，而单模态系统在信息缺失时往往直接失效。从应用层面看，本课题的成果可直接服务于舆情监控系统的情感预警、智能客服的满意度评估以及人机交互中的情绪感知模块，具有切实的工程落地价值。

2 国内外研究现状

多模态情感分析并非一蹴而就，其技术演进经历了从简单到复杂、再到务实回归的曲折历程。早期研究者受限于模型表达能力，普遍采用 **Early Fusion** 策略：将各模态的手工特征（如文本的词袋向量、视觉的 HOG 描述子、声学的 MFCC 系数）直接拼接成高维向量，送入 SVM 或逻辑回归进行分类。这一方法实现简单，却几乎完全忽略了模态间的交互关系。随后，研究者开始探索更复杂的融合机制：Zadeh 等人提出的 **Tensor Fusion Network** 通过外积运算捕捉模态间的高阶交互，**Memory Fusion Network** 则引入了记忆单元建模时序依赖。然而，这些张量方法在提升性能的同时也带来了参数爆炸的副作用，模型规模迅速膨胀。

进入预训练大模型时代后，基于 **Transformer** 的架构成为主流：BERT 及其中文变体在文本语义理解上取得突破，ViT 将图像建模为 Patch 序列打破了 CNN 的垄断，CLIP 更是实现了视觉与语言的跨模态对齐。这些预训练模型为多模态任务提供了强大的特征提取器，但一个常被忽视的事实是——**这些模型的训练与推理对算力的要求极高**。以联合微调 BERT 与 ResNet 为例，其显存占用通常超过 12GB，这在科研机构的 A100 集群上或许不成问题，但对于使用 RTX 3060（6GB 显存）的本科毕设环境而言，几乎意味着无法运行。

与此同时，当前主流研究多基于英文数据集 CMU-MOSI 进行实验，而中文语境存在独特的挑战：四声调的韵律变化对情感判断有显著影响，网络流行语与谐音梗的使用极为普遍，" 阴阳怪气" 的表达方式更是难以被字面语义捕捉。国内虽有学者开始关注这一问题，但高质量的中文多模态数据依然稀缺。清华大学于 2020 年发布的 **CH-SIMS** 数据集是目前为数不多的选择，其样本来源于非受控的影视片段，包含背景杂音、人脸遮挡等真实噪声，这既是挑战，也为模型鲁棒性的验证提供了理想的试验场。

基于上述分析，本课题明确选择**回归 Early Fusion 的务实路线**：通过冻结 BERT 与 ResNet 的骨干参数、仅训练分类头的方式，将显存需求压缩至 6GB 以内；同时采用梯度累积策略模拟大批量训练。这一工程导向的设计并

非对 SOTA 的盲目追逐，而是在有限资源约束下追求**性价比最高的可落地方案**——这恰恰是本科毕业设计应有的务实态度。

3 数据集介绍

3.1 CH-SIMS 数据集选用分析

在多模态情感分析领域，目前主流的数据集如 CMU-MOSI 和 CMU-MOSEI 均基于英语语料，而高质量的中文多模态情感数据集相对匮乏。直到 2020 年，清华大学发布了 **CH-SIMS** (Chinese Single- and Multi-modal Sentiment Analysis) 数据集，才为该领域提供了重要的中文基准。

CH-SIMS 的样本来源于**电影、电视剧、综艺节目**等非受控环境，共计 2,281 个视频片段。与实验室受控采集的数据相比，这种自然场景下的数据虽然能够更真实地反映人类情感表达，但也引入了复杂的噪声，如背景音乐干扰、非正面人脸、光照变化等。从工程实践的角度来看，这些挑战恰好能够检验模型在实际应用中的鲁棒性。

该数据集将样本划分为训练集 1,368 例、验证集 456 例与测试集 457 例。值得注意的是，CH-SIMS 提供了**三模态独立标注**，即文本、视觉、声学三个模态分别拥有独立的情感标签。这一特性使得我们能够在消融实验中精确量化各模态的贡献，并深入研究模态间的不一致性（例如文本积极但语气消极的”模态冲突”现象）。

3.2 类别不平衡问题的处理

通过对训练集分布的统计分析，我们发现正、中、负三类样本的分布存在显著的不平衡现象，负面与中性样本的数量多于正面样本。若直接使用标准的交叉熵损失函数进行训练，模型可能会倾向于预测多数类以降低整体损失，从而忽略少数类样本的学习。

为了解决这一问题，我们在损失函数中引入了**类别权重（Class Weights）**策略。具体而言，为样本量较少的类别分配更高的权重系数，增加其在损失计算中的比重，从而引导模型均衡学习各类情感特征。这一策略在不增加并造成过拟合风险的前提下，有效地缓解了类别不平衡带来的负面影响。

3.3 数据预处理与噪声处理

鉴于 CH-SIMS 数据集的非受控特性，我们在训练前构建了一套完整的数据清洗管线，以消除噪声对模型性能的影响。面对原始数据中存在的无效帧问题，MTCNN 算法被部署用于人脸检测与过滤；在处理过程中，约有 8% 的视频帧因背对镜头或遮挡导致无法检测到有效人脸，针对这些样本，我们选择将其直接剔除以保证输入数据质量。针对剪辑转场常见的全黑画面，我们引入了基于像素方差的黑帧过滤机制，自动识别并移除低方差的无效帧，防止其干扰特征提取。在此基础上，声学模态的处理聚焦于静音片段的干扰，通过计算均方根（RMS）能量筛选出低能量片段，并在后续训练中对其进行掩码（Mask）处理。上述清洗步骤完成后，利用 CH-SIMS 提供的词级对齐信息，视频帧序列与音频片段被精确对齐至文本的词级粒度，从而确保了多模态特征在时序上的严格一致性。

4 核心技术路线

4.1 架构设计与资源约束

本研究在设计模型架构时，充分考虑了硬件环境的限制，即需要在 **RTX 3060 Laptop（6GB 显存）**上完成模型的训练与推理。这意味着高计算复杂度和高显存占用的模型（如大规模张量融合网络、端到端微调的跨模态 Transformer）难以直接部署。

鉴于此，我们采用了工程上更为可行的**三塔架构（Three-Tower Architecture）结合早期融合（Early Fusion）**方案。在该架构中，文本、视觉、声学三个模态分别通过独立的编码器提取特征，随后将特征向量进行拼接，并送入共享的分类器进行预测。这种设计在保证基本多模态交互能力的同时，显著降低了显存占用和计算开销，符合本科毕业设计的实际硬件条件。

4.2 文本塔设计

在文本特征提取方面，我们利用了 HuggingFace 提供的 `bert-base-chinese` 预训练模型。该模型基于大规模中文语料训练，具备强大的语义理解能力。

具体处理流程如下：首先利用 BertTokenizer 对输入文本进行分词，并添加 [CLS] 与 [SEP] 标记；随后将序列输入 BERT 的 12 层 Transformer 编码器。最终，我们选取最后一层的 **[CLS] 向量**（768 维）作为句子的全局语义表征。由于 [CLS] 标记在预训练阶段已被专门训练用于聚合序列信息，因此直接作为句子表示是标准且有效的做法。

4.3 视觉塔设计与时序聚合

视觉模态的处理对显存资源的消耗最为显著。ResNet-50 提取的每一帧特征维度为 2048。如果采用 LSTM 或 GRU 等循环神经网络对帧序列进行时序建模，其梯度反向传播过程中的显存占用极易超出 6GB 的限制。

为解决这一问题，我们采用了**时序均值池化（Temporal Mean Pooling）**策略。对于一个包含 N 帧的视频片段，我们将所有帧的特征向量进行平均：

$$\mathbf{V} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i, \quad \mathbf{v}_i \in R^{2048} \tag{1}$$

这一操作将变长的视频序列压缩为固定的 2048 维向量，虽然在一定程度上损失了细粒度的时序动态信息，但有效地控制了显存占用和计算复杂度。

在特征提取层面，我们选用在 ImageNet 上预训练的 **ResNet-50** 作为骨干网络。输入图像经过预处理归一化为 224×224 尺寸，移除全连接层后的 AvgPool 输出即为帧级特征。

4.4 声学塔设计

声学模态的处理采用 **OpenSMILE** 工具包提取底层声学特征，主要包含 13 维 MFCC（及其一阶、二阶差分）和 12 维 Chroma 特征，共计 51 维。这些特征能够有效捕捉语音中的韵律、音高和能量变化。

为了处理变长的声学特征序列，我们设计了一个轻量级的**一维卷积网络（Conv1D）**。该网络通过卷积操作捕捉局部的时序模式，并通过均值池化将其聚合为 256 维的固定长度向量。这一设计在保留关键声学线索的同时，保持了参数量的轻量化。

4.5 融合策略

在获得文本 **T**、视觉 **V** 和声学 **A** 三个特征向量后，我们采用**直接拼接（Concatenation）**的方式进行融合：

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

拼接后的向量 **F** 经过 Dropout ($p = 0.3$) 正则化处理，输入全连接层映射至情感类别空间，最后通过 Softmax 函数输出预测概率。

尽管拼接策略相对简单，但在计算资源受限的场景下，它是平衡性能与效率的有效选择。

4.6 显存优化策略

为了在 6GB 显存环境下顺利训练 BERT 和 ResNet 主干网络，我们制定了一套层层递进的资源适配方案。面对 BERT 庞大的参数量，我们并未选择全量微调，而是实施了**冻结骨干网络（Freeze Backbone）**策略，即在训练初期锁定 BERT 和 ResNet 的参数，仅开放分类头和声学网络的梯度更新，待模型收敛后再解冻顶层微调，从而大幅降低显存峰值。在此基础上，为了突破物理显存对 Batch Size 的限制，**梯度累积（Gradient Accumulation）**技术被应用于训练循环中；该技术通过累积多次前向传播的梯度（例如物理 Batch Size 为 4，累积 8 次）再执行一次参数更新，等效实现了大批量训练的效果。同时，为了进一步压榨硬件性能，**混合精度训练（FP16）**借助 PyTorch

的 AMP 模块被引入全流程，通过在计算中使用半精度浮点数，不仅显著降低了显存占用，更在保证收敛精度的前提下大幅提升了训练速度。

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 1: 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

本项目的核心实验预期在于验证模态间的互补性。具体而言，我们预计全模态模型（Full Model）将在 Accuracy 与 Weighted F1 指标上全面超越所有单模态基线，以此证明多源信息的融合能够有效消除单一视角的歧义。同时，我们推测文本模态在单模态测试中将表现出最强的判别力，验证其在情感分析中的主导地位；而双模态组合（如视觉 + 声学）的性能表现，则将揭示在缺乏语义信息的情况下，单纯依靠非语言线索（表情与语调）进行情感诊断的可行性边界。

5.3 评估指标

- **Accuracy (准确率)**：

$$\text{Accuracy} = \frac{\text{正确预测数}}{\text{总样本数}} \tag{3}$$

  衡量模型的整体分类正确率。

- **Weighted F1 (加权 F1 值)**：

$$\text{Weighted F1} = \sum_{c=1}^{C} \frac{n_c}{N} \cdot F1_c \tag{4}$$

  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

5.5 预期产出

本项目最终将交付一套完整的 PyTorch 工程源码，包含数据清洗管线、三塔模型实现以及训练评估脚本，并配以详细的文档注释以确保可复现性。此外，为了直观展示模型效果，我们将构建一个基于 Gradio 的交互式演示界面，支持用户上传视频片段并实时可视化情感预测结果与各模态的置信度分布。最终的研究成果将汇总为一篇符合学术规范的毕业论文，详细记录消融实验的数据对比、模态互补性分析以及模型在非平衡数据上的表现，为中文多模态情感分析提供有价值的参考。

表 2: 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)

6 进度安排

本毕业设计计划于 2026 年 1 月至 5 月期间完成，整体进度安排紧凑且目标明确。

第一阶段为**数据准备与预处理期**（1 月至 2 月）。作为项目的基石，本阶段的核心任务是确保数据处理流程的稳健性。这包括：实现基于 MTCNN 的人脸检测与关键点定位脚本，确保在 CH-SIMS 数据集上能够稳定运行；利用 OpenSMILE 批量提取声学特征并进行标准化处理；解析官方对其文件，完成三模态数据的时序对齐。该阶段的完成标志是建立一套可复现的、无错误的预处理管线。

第二阶段为**模型构建与基线验证期**（3 月）。工作重点将转向模型的具体实现。我们将首先构建单模态基线（特别是文本模态），验证数据流的可行性。随后，逐步接入视觉和声学模块，构建完整的三塔架构。在此期间，显存优化策略（如冻结骨干、梯度累积、混合精度训练）将是调试的重点，以确保模型能在硬件限制下正常训练。本阶段的目标是实现三模态模型的端到端训练，并观察到验证集损失的有效下降。

第三阶段为**实验与优化期**（4 月）。此阶段将集中进行系统的消融实验，验证各模态及融合策略的有效性。我们将记录并分析各实验组的 Accuracy 和 Weighted F1 指标，绘制性能对比图表。此外，如有余力，将尝试微调 BERT 更多层数等策略以探索性能上限。

第四阶段为**论文撰写与答辩准备期**（5 月）。根据实验数据和研究成果，撰写学位论文，内容涵盖研究背景、方法论、实验设计及结果分析等。同时，制作答辩演示文稿（PPT）及可视化的系统演示 Demo，确保能够直观展示研究成果。最后，对项目代码进行整理和文档化，以便于后续的存档与交流。

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.