当代人工智能 实验五
多模态情感分类

周欣怡 10233330402 新闻双

GitHub 仓库链接：https://github.com/Chloris0511/Final_Project

一、多模态情感分析模型设计与说明

（一）模型设计动机

情感分析任务中，单一模态往往难以完整表达用户的真实情绪状态。文本信息能够表达主观语义，但在表达情绪强度、场景氛围等方面存在局限；图像信息能够提供视觉线索，但缺乏明确的语义解释能力。而多模态情感分析模型将文本与图像信息进行联合建模，能充分利用不同模态之间的互补性。

（二）模型结构设计

模型整体由三部分组成：

1. 文本编码器：采用预训练语言模型 BERT，对文本进行上下文语义建模；
2. 图像编码器：采用 ResNet-18 提取图像高层视觉特征；
3. 融合与分类模块：通过跨模态注意力机制（Cross-Attention）对文本与图像特征进行交互建模，并最终完成情感分类。

在融合阶段，模型以文本特征作为查询（Query），图像特征作为键和值（Key / Value），通过注意力机制引导模型关注与文本语义最相关的图像信息。

二、模型亮点

在多模态情感分析任务中，常见的融合方式主要包括特征拼接或加权平均。尽管这类方法实现简单，但其本质是对不同模态特征进行静态融合，难以刻画文本与图像之间复杂且动态的语义关联关系。所以我们在多模态融合阶段引入了跨模态注意力机制（Cross-Attention），以增强模型对不同模态间关键信息的选

择与对齐能力。

具体而言，模型首先分别对文本与图像进行特征编码，得到高层语义表示。在此基础上，实验将文本特征作为查询向量，图像特征作为键和值，通过注意力计算机制显式建模文本语义与图像内容之间的相关性。该设计使模型能够在不同样本中，根据文本内容的变化动态关注图像中更具情感判别力的区域特征，而非对所有图像信息进行等权处理。

与传统 Late Fusion 方法相比，跨模态注意力机制在融合阶段引入了显式的信息交互过程，使不同模态不再是独立决策后的简单组合，而是在特征层面实现深度协同。一方面，该机制能够显式学习模态之间的对齐关系，避免无关或噪声信息对最终分类结果产生干扰；另一方面，注意力权重的引入使融合过程具备一定的可解释性，有助于分析模型在不同样本中更依赖哪一类模态信息。

三、实验过程中的问题与解决方法

（一）数据划分与路径组织问题

在初始实验阶段，由于数据集中同一条样本的文本与图像文件位于同一目录下，而训练与验证集又需要分别管理，导致在数据加载过程中频繁出现文件路径错误。在路径组织上一度出现混乱，找不到需要的文件。

针对该问题，实验中采用了以下解决方案：

在 Dataset 初始化阶段通过解析传入的 split_csv 文件名自动判断当前数据集所属的划分（train / val），并据此动态设置文本与图像的根目录路径，从而确保 Dataset 在不同阶段均能正确定位对应的数据文件。

（二）文本编码与字符集问题

在文本清洗阶段，部分原始文本文件包含非 UTF-8 编码字符，直接读取时会触发 UnicodeDecodeError，导致预处理脚本中断。

对此，实验在文本读取阶段统一使用 errors="ignore"方式进行编码容错处理，确保文本内容能够被稳定读取并进入后续模型处理流程。

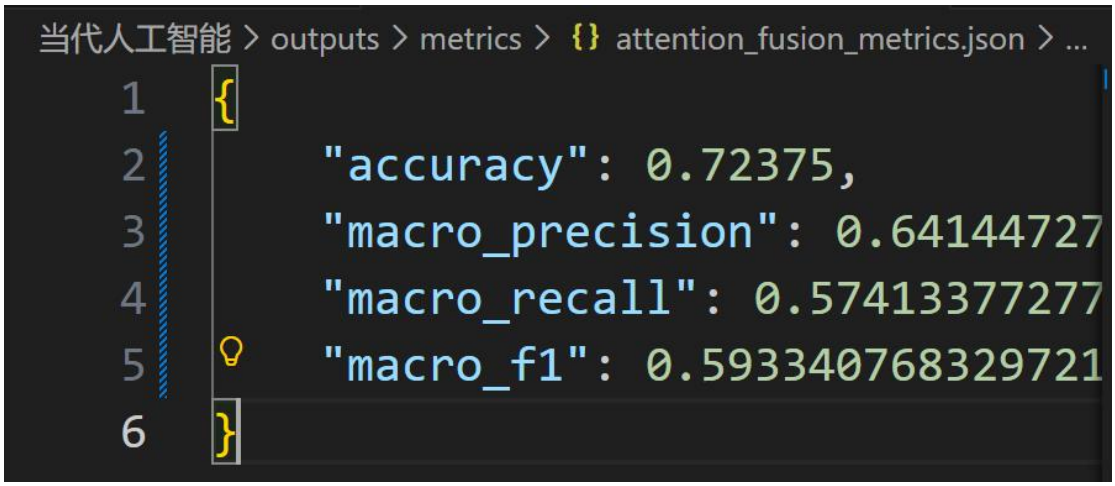
（三）标签类型不匹配问题

在模型训练阶段，最初直接尝试将 csv 文件中的情感标签（如 positive、neutral、negative）强制转换为整数，导致 ValueError 异常。

最终的解决方案是在 Dataset 类中显式定义标签映射表（label map），将字符串形式的情感标签统一映射为离散整数标签，从而保证模型训练与损失计算的正确性与一致性。

四、实验结果分析与结论

（一）验证集实验结果



attention_fusion_metrics.json 截图

在验证集上，多模态注意力融合模型整体表现较为稳定，取得了较为理想的分类效果。从整体指标来看，模型在验证集上的准确率达到 0.7238， Macro-F1 值为 0.5933，Macro-Precision 和 Macro-Recall 分别为 0.6414 和 0.5741。在类别分布相对不均衡的情感分类任务中，Macro-F1 能够较为客观地反映模型在各类别上的综合性能，因此该结果表明模型在不同情感类别上的判别能力整体可接受。

（二）消融实验结果分析

消融实验分别评估了以下三种模型设置：

1. Text-only 模型：仅使用文本特征进行情感预测；
2. Image-only 模型：仅使用图像特征进行情感预测；
3. Multimodal 模型：融合文本与图像特征并引入注意力机制。

	A	B	C	D
1	model	ablation	val_acc	val_macro_f1
2	attention_	multimoda	0.7125	0.5801
3	attention_	multimoda	0.705	0.5849
4	attention_	text	0.6975	0.5678
5	attention_	image	0.625	0.5115

test_predictions.csv 截图

从结果可以看出，Text-only 模型在验证集上取得了 0.6975 的准确率和 0.5678 的 Macro-F1，整体表现优于 Image-only 模型，说明在该情感分析任务中，文本信息仍然是情感判别的主要信息来源。

多模态注意力融合模型在两项核心指标上均优于单模态模型，其验证集准确率达到 0.7125，Macro-F1 达到 0.5801。与 Text-only 模型相比，多模态模型在准确率和 Macro-F1 上均有所提升，说明引入图像信息并通过注意力机制进行跨模态融合，能够在一定程度上补充文本特征，提升模型对情感的整体判别能力。

综上所述，消融实验结果验证了多模态融合策略的有效性：文本信息在情感分析中起主导作用，而视觉信息作为辅助模态，通过注意力机制参与特征融合后，能够进一步增强模型性能，体现了多模态情感分析方法的实际价值。