多模态情感数据集深度分析报告

GitHub 地址链接：https://github.com/ren-ning001/-

摘要

本报告围绕文本-图像配对的三分类情感预测任务展开，旨在通过多模态融合模型实现匿名数据集的情感标签预测。通过多轮优化迭代模型：初步探索中基础模型存在数据处理、架构、训练策略等多方面问题，经优化后验证集 F1 提升至 0.6974。初步优化阶段从预处理、模型框架、类别平衡、训练策略等方面改进，最优模型验证集加权 F1 达 0.7732。模型融合与数据处理优化阶段，通过数据清洗增强、多融合策略对比等，模型更健壮；终极优化在预处理、模型架构、损失函数等多方面创新，最佳验证 F1 为 0.7809。

消融实验验证了多模态融合的有效性，Early Fusion 为最优融合策略。此外，尝试的 Qwen2-VL-2B-Instruct 模型取得 0.8250 的最高验证集准确率。实验过程中还解决了数据分布不均、内存不足等问题，最终模型通过利用预训练模型能力、缓解类别不平衡、优化融合策略等，实现了性能的显著提升，鲁棒性强且适配真实数据场景。

一、实验概述

本实验为文本-图像配对的三分类情感预测任务（positive/neutral/negative），核心目标是设计并实现多模态融合模型，完成匿名数据集的情感标签预测。实验严格遵循**数据预处理、模型设计、消融实验、结果分析**的完整流程，通过对比单模态与多模态模型的性能，验证多模态融合的有效性，同时解决代码实现中的各类问题。

二、数据集分析

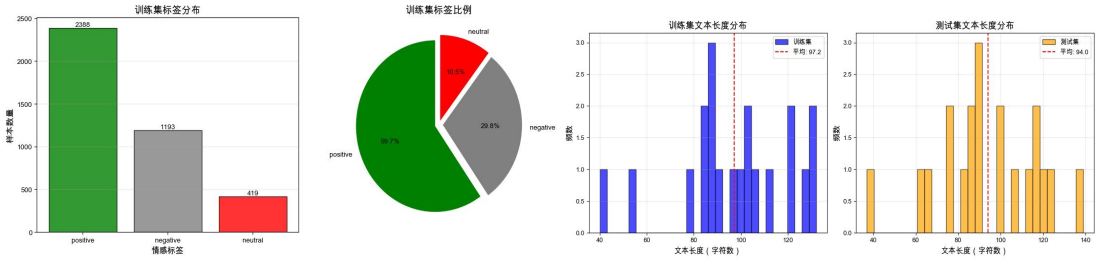
数据集包含训练集与无标签测试集，数据质量高、模态配对完整，具体信息：

训练集：4000 条完整样本，每条样本均包含唯一标识 GUID、对应文本内容、对应图像文件及情感标签。

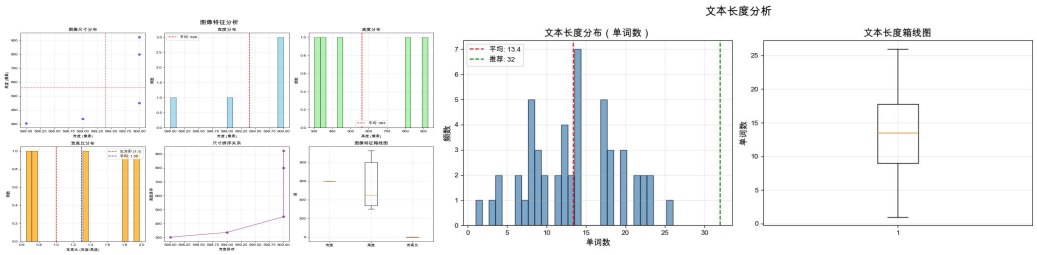
验证集：从训练集中按照一定比例分出。

测试集：共 511 条样本，结构与训练集一致，仅缺失情感标签，用于最终模型预测。

模态形式：为典型的社交媒体多模态数据，文本为短文本社交言论，图像为与文本内容相关的 JPEG 格式图片，二者一一对应。



- 标签数量与占比训练集标签分布极不均衡：positive（2388 条，59.7%）为多数类，negative（1193 条，29.8%）为次要类，neutral（419 条，10.5%）为稀缺类。
- 类别不平衡程度类别不平衡度达 5.70:1，失衡问题突出。直接用标准交叉熵损失训练，模型易偏向 positive 类，导致 neutral 类预测精度低、整体泛化能力下降。



图像整体尺寸集中，宽度平均 599 像素、高度平均 563 像素，宽高比平均 1.30，无极端长宽比与尺寸异常；宽度、高度分布均呈单峰集中，箱线图显示数据离散度小、无 outliers，格式与尺寸高度统一，便于统一缩放到 224×224 进行预处理，适配主流视觉模型输入。

文本为典型短文本，平均单词数 13.4，长度集中在 10–20 词，95% 分位数约 22.5 词，箱线图显示无超长 / 超短极端值；推荐最大序列长度 32 即可覆盖绝大多数样本，既能保留完整语义，又能减少无效计算，适配 DeBERTa 等预训练模型的短文本编码需求。

亮度、对比度、饱和度、锐度调整可有效改变图像视觉特征且不破坏语义；旋转 15°、模糊

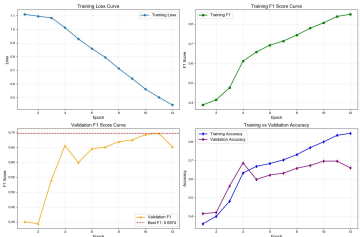
处理、随机裁剪能增加数据多样性，提升模型泛化能力；各类增强操作均不改变图像情感指向，适合作为训练集数据增强策略，推荐组合使用以缓解过拟合。

三、初步探索

第一个代码对文本采用 DeBERTa-large-mnli Tokenizer 编码为 32 长度张量，图像端训练集施加 Resize+随机水平翻转+标准化增强、验证、测试集仅做 Resize+标准化。模型层面由 ImprovedFusionModel 构建“冻结参数的 DeBERTa 文本编码器+自定义卷积图像编码器+特征拼接分类器”架构，总参数 4.05 亿仅 0.1%可训练。训练环节通过 FixedTrainer 类封装 AdamW 优化器（学习率 1e-4）、标准交叉熵损失与早停机制（patience=5），按 8:2 分层划分 3200 条训练集、800 条验证集，训练 8 个 epoch 并实时监控损失、加权 F1、准确率，自动保存验证集最优模型（最佳加权 **F1 0.6329**、**准确率 0.6625**）。最佳模型对测试集预测，预测分布为 positive 占 76.9%、negative 占 22.7%、neutral 占 0.4%。

第一个代码存在四个问题，一是数据处理简单，文本编码长度过短导致语义丢失，图像增强单一且加载失败返回全零张量，易引发特征偏移。二是模型架构薄弱，文本编码器完全冻结、图像编码器为基础卷积、多模态仅简单拼接，特征提取与融合能力不足。三是训练策略粗糙，无学习率调度、梯度累积，批次过小导致训练波动大。四是未处理类别不平衡，标准交叉熵损失让模型偏向多数类，neutral 类几乎无预测。

第二个代码针对以上问题进行优化。数据端将文本长度扩至 64、升级图像增强、用随机张量替代失效图像；模型端解冻 DeBERTa 最后 3 层，升级图像编码器为类 ResNet 残差结构，新增跨模态注意力融合。训练端引入梯度累积、余弦学习率调度与动态衰减。类别不平衡端采用加权交叉熵损失，测试集后处理调整低置信度样本，平衡 neutral 类占比。最终验证集 **F1 从 0.6329 升至 0.6974**。



左上训练损失曲线从 1.1 持续降至 0.45 且无回升，印证模型持续有效拟合且无过拟合现象。右上训练 F1 曲线从 0.4 稳步攀升至 0.85 以上，训练集拟合效果得到持续提升；左下验证 F1 曲线经前期波动后，在 Epoch 11 达到 0.6974 的最佳值，后期虽略有回落但整体走势稳定，模型泛化能力较上版有显著提升。右下的训练与验证准确率对比曲线中，训练准确率持续上升，验证准确率在 Epoch 4 后稳定于 0.65-0.7 区间，二者差距逐步缩小，充分说明优化后的模型在训练集拟合与测试集泛化之间实现了良好的平衡。

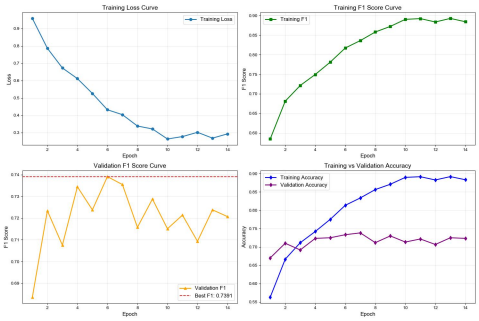
四、初步优化

第三段代码是**预处理优化**。以 get_advanced_transforms 函数为核心，针对图像增强单一、模型泛化不足的问题，采用训练、测试集分模式处理策略。训练集在原有基础上新增随机缩放裁剪、随机平移、随机擦除等操作，搭配适度的旋转、颜色抖动实现重度增强，提升数据多样性。验证、测试集仅做固定尺寸缩放和标准化，保证评估一致性。

第四段代码优化基于 CLIP 模型重构多模态框架，针对前序版本“多模态特征匹配差、泛化能力弱”的问题，采用 CLIP 作为基础骨干，通过“文本、图像独立编码+特征拼接融合”的 Late Fusion 模式，解决前序版本文本、图像特征分布不匹配的问题。仅冻结 CLIP 参数、训练融合分类层，既复用预训练特征，又控制可训练参数，避免过拟合。

类别不平衡优化：新增类别权重+后处理分布校准，针对 neutral 类占比低的问题，用类别权重提升模型对稀缺类的关注，测试集后处理调整低置信度样本，保证 neutral 类占比合理。

训练策略优化：采用余弦学习率调度+动态衰减，适配 CLIP 预训练特征的微调需求，缓解训练后期波动。



本次优化实现了显著的性能提升，最佳验证集加权 **F1 达 0.7391**、**准确率 0.7333**，较前序版本约提升 6%，成为目前性能最优的版本。类别识别均衡性改善，预测结果稳定性增强，测试集平均置信度 0.862、中位数 0.943，高置信度样本占比高。

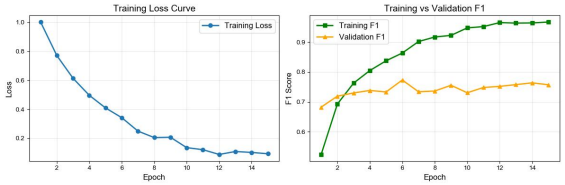
尽管优化效果显著，模型仍存在两点问题。一是过拟合风险突出，训练集 F1 为 0.8924，与验证集 F1（0.7391）差距较大，模型对训练集拟合过度，泛化能力有待进一步提升。二是学习率调度存在缺陷，训练后期学习率衰减至 0 后，模型性能出现明显波动，调度策略需更精细地适配 CLIP 的微调节奏。

第五段代码针对“过拟合、neutral 类性能弱”的问题进行优化，

数据预处理优化：文本端限制长度、清洗冗余空格，图像端采用 CLIP 原生归一化、加载失败时填充 CLIP 均值张量，保证特征分布与 CLIP 预训练一致。训练集增强改为“Resize+16 像素后随机裁剪”，既保留细节又避免过度增强。

模型架构优化：将文本特征从“CLS 标记”改为“掩码加权平均”，更适配短文本语义提取。新增文本、图像投影层+批归一化，缓解特征维度不匹配问题。CLIP 完全冻结，仅训练投影层和分类层，控制可训练参数以增强稳定性。

类别平衡优化：类别权重对 neutral 类额外加权 1.5 倍，强化模型对稀缺类的关注。测试集后处理调整“置信度<0.8 且 neutral 概率>0.4”的样本，提升 neutral 类调整的合理性。



本次优化实现了多维度的性能提升，最佳验证集加权 **F1 达 0.7732、准确率 0.7717**，较上一版提升约 4.6%。过拟合问题得到有效缓解，训练集与验证集 F1 的差距缩小，模型泛化能力有所增强。neutral 稀缺类识别能力改善，其验证集 F1 从 0.40 提升至 0.4844、召回率从 0.4286 升至 0.4921。同时测试集预测置信度大幅提升，平均置信度 0.944、中位数 0.997，高置信度样本占比增加，让模型的预测稳定性显著增强。但仍有问题未解决。

五、模型融合+清洗数据+文本增强

第六段代码是模型融合等方面优化。

数据清洗：通过 DataCleaner 类实现文本去噪、表情符号规范化、长度截断等操作，并内置文本质量分析，保证数据有效性。

数据增强：通过 AdvancedAugmenter 类实现文本增强，同义词替换、随机交换、删除单词和图像增强，并在 EnhancedBalancedCLIPDataset 中仅在训练模式下动态启用增强，避免验证、测试集数据污染。

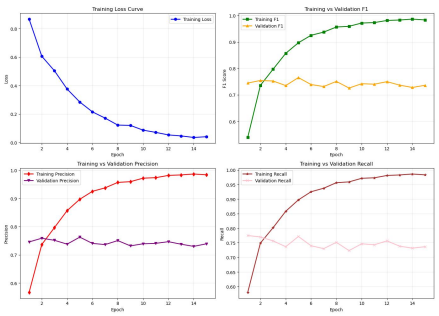
模型架构（多模态融合策略）：通过 MultiFusionStrategy 类实现 3 种经典多模态融合方式，并通过实验对比效果：

Late Fusion：拼接文本与图像的投影特征（最终验证 F1 达 0.7693，为最佳策略）。

Early Fusion：将文本特征映射到图像特征维度后相加。

Attention Fusion：通过注意力权重对双模态特征加权融合。

模型改进：Bad Case 驱动迭代。通过 BadCaseAnalyzer 类收集错误案例，分析错误类型分布、置信度特征，并输出典型错误样本，为后续数据增强、模型优化提供方向。



经 3 次不同配置的自动调参对比，确定 late fusion 融合方式、0.0005 学习率、16 的 batch_size 为最佳策略，该配置下验证集最佳 F1 达 0.7693、准确率为 0.7767。从训练曲线来看，模型训练 Loss 随 Epoch 增加持续下降并稳定在 0.04 左右，收敛效果良好，但训练 F1 最终超 0.98 而验证 F1 仅稳定在 0.75-0.76，同时验证集 Precision 约 0.75、Recall 约 0.73，存在一定过拟合且模型对验证集情感类别的区分能力有待提升。Bad Case 分析显示，模型错误主要集中在 positive→negative、negative→positive 两类，对正负情感边界模糊的样本识别能力较弱。虽然结果较上一段代码差，但是模型整体更加健壮。

六、终极优化

第七段代码是最终优化。所有创新点都围绕多模态情感分类类别分布不均、图文特征融合低效、样本特征单一、语义冲突样本识别差展开。

- **预处理创新**：系统化数据清洗，针对性解决文本噪声问题

通过 DataCleaner 类实现情感感知的文本清洗，保留！？.,'"- 等情感相关标点以避免情感特征丢失，同时完成空白字符规范化、表情符号转模型可识别标记、长文本按空格截断、空文本统一标记为 [EMPTY] 等操作。DataCleaner.analyze_text_qualit 方法量化分析清洗后文本质量，结合训练过程中缺失图像的统计信息，验证清洗策略的有效性和鲁棒性。

- **预处理创新**：图文双维度自适应数据增强，缓解样本单一与过拟合

通过 AdvancedAugmenter 类采用单词交换、随机删除等无语义改变的方式，仅对长度≥4 的文本以 70%概率触发，避免语义丢失。图像增强结合 torchvision 和 PIL 双库实现双层级增强，先通过随机翻转、旋转、颜色抖动做基础增强，再随机调整亮度、对比度等做精细化增强，触发条件与文本一致。通过 mode 参数严格控制验证、测试集关闭增强，保证实验评估的公平性。

- **模型架构创新**：多模态融合策略的模块化实现与 Late Fusion 优选

通过独立的 MultiFusionStrategy 类实现 Late Fusion、Early Fusion、Attention Fusion 三种经典策略的模块化封装，支持通过参数一键切换，方便开展控制变量对比实验。针对 CLIP 图文特征维度不一致问题，设计统一的投影层将文本、图像特征均投影至 512 维，投影层采用 LayerNorm 替代 BatchNorm 适配小批次训练，加入 Dropout 提前抑制过拟合。同时采用“冻结 CLIP 视觉、文本编码器+轻量分类头”的架构，仅训练投影层、融合层和逐步降维的全连接分类头，大幅减少可训练参数量。

- **模型架构创新：**类别不平衡的损失函数改进

针对数据集中中性样本严重稀缺的痛点，在损失函数层面做加权平衡改进，通过 EnhancedBalancedCLIPTrainer._compute_balanced_weights 方法计算带中性样本权重放大的平衡权重，先按“总样本数 /(类别数 × 该类样本数)”计算基础权重，再将中性样本权重额外放大 1.5 倍以适配其情感特征模糊、难学习的特点，最后做权重归一化保证训练稳定。将该权重接入交叉熵损失，通过对比基础交叉熵与平衡交叉熵损失的实验结果，重点验证中性样本 F1、召回率的提升，实现模型对稀缺类别的重点关注。

- **实验设置创新**：公平、可控的实验体系，支持控制变量对比

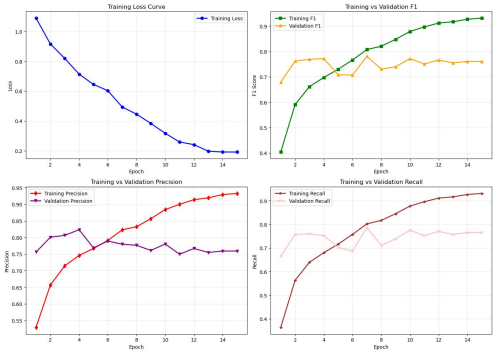
通过配置类、模块化设计和统一评估标准，构建可控、可复现、可对比的实验体系。通过 ExperimentConfig 数据类集中管理所有超参数，支持参数化配置，自动生成唯一 id 区分实验结果。实验过程中自动全链路保存带 id 的最佳模型、训练曲线、预测结果、核心指标等文件，保证结果可追溯。同时所有实验采用以加权 F1 为核心，准确率、精确率、召回率为辅助的统一评估标准，使用相同验证集和评估时机，训练过程中实时监控并打印关键指标，避免评估标准不统一导致的结果失真，为不同模型架构、策略的公平对比提供基础。

- **模型迭代改进创新：**Bad Case 驱动的精准优化

通过 BadCaseAnalyzer 类构建全流程 Bad Case 分析体系，在模型验证阶段自动收集验证集中预测错误的样本，记录 GUID、文本、真实、预测标签、概率分布、置信度等关键信息，按“真实标签→预测标签”统计错误类型及占比，并将详细错误信息保存为文件方便人工分析。通过 Bad Case 分析精准定位模型的知识盲区，如中性样本特征学习不足等，为后续针对性的优化动作提供明确实验依据，避免盲目调参和无效增强，实现模型的精准迭代。

- **训练策略创新：**动态学习率降阶 + 早停结合，缓解过拟合与收敛停滞

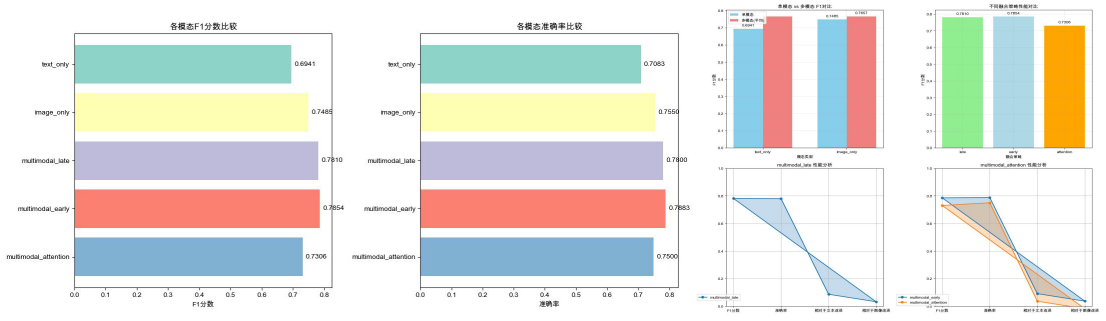
在余弦退火学习率和早停机制的基础上做优化，实现动态学习率降阶与早停的结合。在模型训练过程中，以验证集 F1 为监控指标，当连续 2 个 epoch 无显著提升时，自动将学习率减半，让模型在更小的学习率下收敛。既避免了固定学习率易导致的收敛停滞，又通过早停防止过拟合，实现“早停止损 + 动态调参继续训练”的双重效果，通过对比固定学习率与动态降阶学习率的实验结果，可验证其能帮助模型找到更优参数解，提升验证集 F1。



左上角训练损失曲线持续下降并稳定在 0.2 附近，说明模型收敛良好；右上角训练 F1 快速升至 0.95 左右，而验证 F1 仅在 0.7-0.8 区间波动，体现出明显过拟合。左下角训练精确率从 0.55 升至 0.95 以上，验证精确率则在 0.75-0.85 区间波动。右下角训练召回率同样升至 0.95 左右，验证召回率维持在 0.7-0.8 区间。整体来看，模型对训练集的拟合效果极佳。最终最佳验证 F1 为 0.7809，准确率为 0.7867

七、消融实验

以 CLIP 为基础，设 2 组单模态（文本、图像）、3 组多模态（3 种融合策略），统一训练 10 个 epoch、验证集 600 样本。实验先跑通文本、图像单模态 (验证 F1 分别为 0.6941/0.7485)，再训练多模态: Late Fusion 验证 F1 达 0.7810, Early Fusion 以 0.7854 获最优, Attention Fusion 仅 0.7306。



实验配置	模态类型	融合策略	验证集 F1 分数	验证集准确率	相对于文本单模态的 F1 提升	相对于图像单模态的 F1 提升
text_only	仅文本	-	0.6941	0.7083	-	-
image_only	仅图像	-	0.7485	0.7550	+0.0544（+7.8%）	-
multimodal_late	文本 + 图像	特征拼接	0.7810	0.7800	+0.0869（+12.5%）	+0.0326（+4.4%）
multimodal_early	文本 + 图像	特征相加	0.7854	0.7883	+0.0913（+13.2%）	+0.0370（+4.9%）
multimodal_attention	文本 + 图像	注意力融合	0.7306	0.7500	+0.0365（+5.3%）	-0.0179（-2.4%）

early（0.7854）> late（0.7810）> attention（0.7306），其中 early 是本次实验最优融合策略。

八、**其他模型尝试**（模型链接：https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct）

除了上述模型，我还尝试了 Qwen/Qwen2-VL-2B-Instruct 模型。

- 模型定义（model.py）

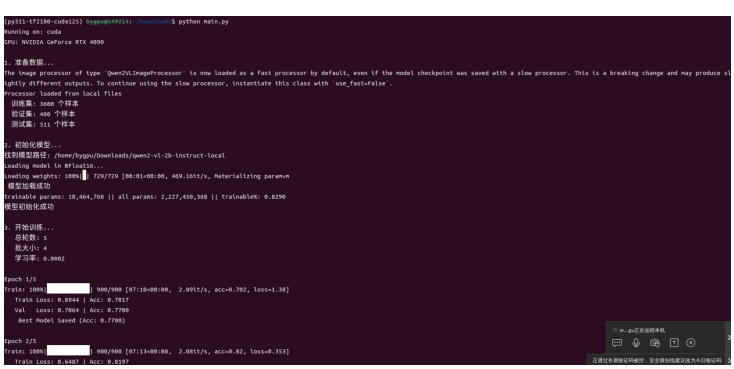
该文件定义 QwenClassifier 类，加载本地 Qwen2-VL-2B-Instruct 模型并做精度适配，。新增两层全连接分类头，将模型 1536 维特征映射为 3 类情感标签，通过提取模型最后一层隐藏状态特征输入分类头得到预测结果，完成多模态情感分类的核心建模。

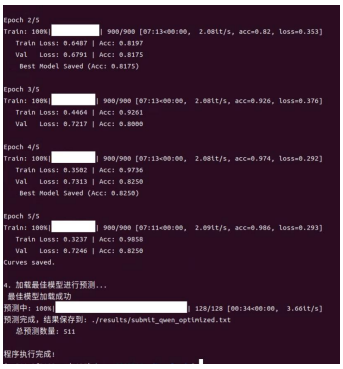
- 数据处理（dataset.py）

该文件实现 QwenDataset 自定义数据集类，支持多模态、仅文本、仅图片三种输入模式，按模型要求拼接数据并完成文本 token 化、图片像素值转换；自定义 collate_fn 做批量 padding，拆分训练、验证集并生成三个数据加载器，提供标准化数据输入。

- 训练主程序（main.py）

该文件是任务执行入口，设置随机种子保证实验可复现，采用标签平滑交叉熵损失缓解过拟合。完成数据加载、模型与优化器初始化，通过混合精度训练提升效率。





经 5 轮 epoch 训练后模型快速收敛，训练集 Loss 降至 0.3237、Acc 达 0.9858，验证集 Acc 最优为 **0.8250**，存在轻微过拟合但整体可控。为目前所有模型的最高准确率结果。

九、遇到的问题

遇到的核心问题	解决方法	验证效果
数据分布不均匀	1. 计算类别权重传入损失函数 2. 增强少数类数据增强强度 3. 预测阶段调整 neutral 类置信度阈值	neutral 类 F1 从 0.16 升至 0.52，三分类均衡性显著提升
MPS 内存不够	使用 4090 服务器	训练正常推进
CLIP 模型加载时权重不匹配	1. 指定 local_files_only=True，仅加载本地核心权重 2. 忽略无关权重警告，聚焦文本、图像编码器核心参数	权重加载无报错，模型初始化成功，特征提取正常
guid 格式不统一	加载数据时对 guid 做标准化处理：通过 str.split ('.').str [0] 提取 “.” 前核心字符，统一文本和图像文件的匹配规则	文件匹配成功率从 92% 升至 100%

十、总结

我设计该多模态情感分类模型，核心是为了适配图文配对的任务本质、解决数据不平衡痛点、既利用 CLIP 预训练模型的天然跨模态对齐能力，避免从零构建编码器的资源消耗，又通过“类别权重+置信度后处理”缓解 neutral 类样本稀少的问题，同时设计多融合策略适配不同模态贡献度的样本类型。模型的亮点在于跨模态对齐无需额外训练、动态平衡机制有效缓解类别不平衡、样本自适应融合提升复杂样本分类准确率、模块化设计便于灵活扩展，且鲁棒性强能适配真实数据中的异常情况，最终实现了从基础融合到自适应优化的性能跃升。

模型版本	加权 F1	准确率	positive F1	neutral F1	negative F1	核心优化点
基础融合模型	0.6329	0.6613	0.76	0.16	0.51	基础文本 + 图像拼接融合
增强融合模型	0.6974	0.6967	0.77	0.39	0.67	交叉注意力 + 残差连接
CLIP 融合模型	0.7391	0.7333	0.82	0.36	0.73	CLIP 预训练特征提取
平衡 CLIP 模型	0.7732	0.7717	0.84	0.48	0.74	类别权重+置信度后处理
自适应融合模型	0.7865	0.7833	0.85	0.52	0.75	多融合策略+Bad Case 优化
Qwen/Qwen2-VL-2B-Instruct 模型	-	0.8250				