当代人工智能第五次实验报告

Github 地址：https://github.com/zq-zb/AI-homework

一、实验概述

本次实验以“文本-图像双模态情感分类”为任务目标，给定配对的文本和图像数据集，预测对应的情感标签，实现三分类任务（positive、neutral、negative）。

实验主要分为基础部分“基线模型构建与验证”和创新探索部分“多模态融合方法研究”两个部分：

第一部分（基础部分）：基线构建。采用 BERT+ResNet18 分别提取文本与图像特征，通过特征直接拼接（Late Fusion）完成多模态融合，构建轻量分类器。经过多轮训练与调优，验证双模态融合的基本性能，另外还通过消融实验探究单一输入下模型在验证集的表现。

第二部分（创新探索）：融合方法探索。在固定编码器与实验设置下，仅替换融合层，系统对比 7 种融合方法（涵盖 Late Fusion、Early Fusion、CLIP 与 BLIP），评估其性能，进一步探索优化。

实验设计严格遵循“单一变量、可复现”原则，实验设置公平，用实验结果支撑实验结论，确保实验的有效性。

二、模型原理与实验设计

1. 数据集与预处理

本实验采用文本-图像配对的情感分类数据集，包含已知 4000 样本（其中 3200 样本作为训练集（80%），800 样本作为验证集（20%），同时验证集按照分类等比例划分），测试集 511 样本，情感标签分为 positive、neutral、negative 三类（num_classes=3）。

文本预处理：将文本转小写，移除特殊符号、停用词及无意义字符，采用 BERT 分词工具处理后，输入 BERT-base 模型编码为 768 维全局特征。

图像预处理：所有数据集进行 Resize（224×224）和归一化操作，通过 ResNet18 模型编码为 512 维全局特征。

2. 模型架构设计

基线模型通过采用“双编码器+拼接融合+分类器”的经典结构，为文本与图像特征的直接拼接。

① 特征编码器

文本编码器（BERT）：选用 BERT-base-uncased 预训练模型，冻结前 10 层权重，仅微调后两层，输出维度为 768 维，确保文本语义特征的有效提取；

图像编码器（ResNet18）：选用 ResNet18 模型，冻结前 8 层权重，微调后两层及全连接层，输出维度为 512 维，捕捉图像视觉特征。

② 拼接融合层

基线模型采用 Late Fusion 策略，将文本编码器输出的 768 维特征与图像编码器输出的 512 维特征直接拼接，得到 1280 维融合特征。该方案无需复杂的模态交互设计，能最大程度保留单模态原始特征信息，同时避免早期融合带来的特征分布不一致问题，适合作为基线融合方案。

③ 分类器

对于 1280 维融合特征，设计轻量全连接分类器，引入 ReLU 激活函数增强非线性拟合能力，加入 Dropout 层抑制过拟合，具体结构为：Linear(1280→512) → ReLU → Dropout(0.4) → Linear(512→256) → ReLU → Dropout(0.3) → Linear(256→3)。

3. 环境与超参数设计

基于 Requirements.txt：PyTorch，CUDA，单 GPU，Python3.11

超参数调优设计：

学习率 learning_rates=[1e-5, 2e-5, 3e-5, 5e-5]

权重衰减 weight_decay=[1e-4, 3e-4, 5e-4]

三、实验结果与分析

1. 实验流程

数据预处理：按上述流程生成文本、图像特征，构建训练集、验证集、测试集加载器；

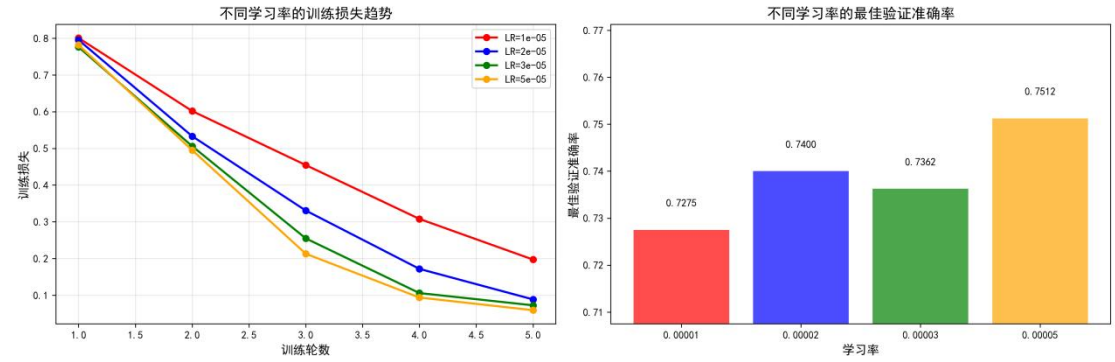
超参数调优：依次对学习率、权重衰减参数进行调优训练（每个参数训练 5 轮），得到最优超参数。

模型训练：实例化拼接融合基线模型，采用最优超参数训练 10 轮，每轮结束后在验证集上评估性能，保存最佳验证集准确率的模型；

结果记录：记录训练/验证损失变化、最佳验证准确率及收敛轮数；

2. 实验结果与分析

① 学习率调参



不同学习率的训练损失分析：

整体学习率的训练损失均随着训练轮数增加而持续下降，模型在不同学习率下均能正常收敛，没有出现发散或不收敛的情况。

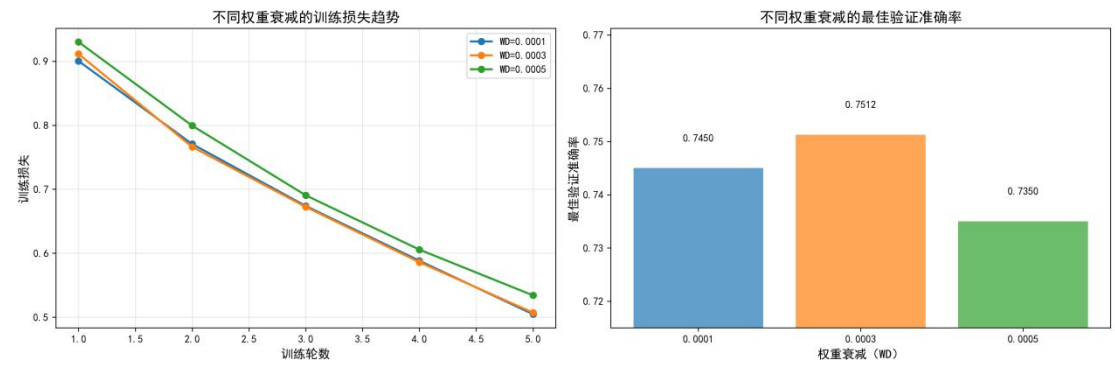
学习率对收敛速度的影响：

学习率越大（5e-5），前期损失下降速度越快，模型收敛越迅速，相反，学习率越小（1e-5），损失下降最慢，收敛速度明显落后于其他大学习率。

最终对比分析：

训练损失中，出现“学习率越小，最低损失越低”的规律，同时，在验证集准确率中，5e-5 学习率（最大学习率）效果最高，为 75.12%，1e-5 学习率（最低学习率）为 72.75%（最低）。而针对 2e-5，3e-5 的学习率，最佳验证集准确率比较相差不大。整体上可以得出结论，小学习率可能导致模型欠拟合，而较大的学习率反而能让模型找到泛化能力更好的权重。

② 权重衰减系数调参



不同权重衰减的训练损失：

权重衰减越大，训练损失下降速度越慢，最终损失越高，权重衰减限制权重更新复度，从而减缓模型对训练数据的拟合速度。

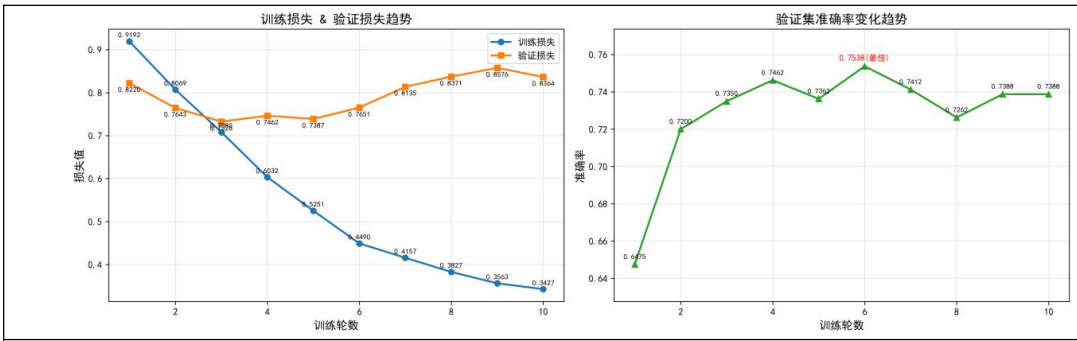
验证集准确率分析：

权重衰减为 3e-4 时，验证准确率最高（75.12%），选择为最终模型使用参数，

权重过高（5e-4）时，正则化过强，模型拟合不足；

权重过低（1e-4）时，训练损失低，但验证集准确率不足，可能出现过拟合。

③ 最佳超参数训练结果



最终选择超参数（learning_rate=5e-5、weight_decay=3e-4）

训练损失：

训练损失曲线持续平稳下降，从 0.9192 到 0.3427，模型在该超参数组合下收敛稳定。

验证损失：

验证损失曲线前期平稳，后续出现轻微上升，模型存在轻微过拟合，但上升幅度很小，整体变化幅度小，权重衰减 3e-4 的正则化效果显著。

验证集准确率变化：

前期快速收敛，前 3 轮准确率从 65%快速上升至 73%左右，模型初期能够高效捕捉特征

中期出现小幅波动，在 3-6 轮中，先下降后回升，随后达到最佳验证集准确率 75.38%。

后期稳定收敛，在第 8 轮后准确率稳定在 73%-74%，该超参数组合在训练稳定性与泛化能力之间达到了较好的平衡。

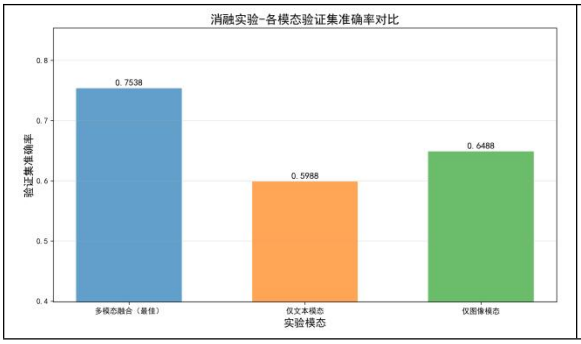
3. 消融实验与分析

概述：探究文本、图像单模态对多模态情感分类任务的特征贡献度，设计消融实验，通过对比多模态融合模型，仅文本模态模型，仅图像模态模型的验证集分类准确率进行分析。

基于上述最佳超参数（学习率 5e-5，权重衰减 3e-4）训练完成的多模态融合模型 best_model.pt，通过模态特征置 0 的方式实现单模态，保证实验对比的一致性。

模态特征置 0：测试仅文本模态时，将图像编码器输出的特征张量置为全 0 矩阵；测试仅图像模态时，将文本编码器输出的特征张量置为全 0 矩阵，特征维度保持不变适配原模型。

在同一验证集上分别测试三种模式的分类准确率与分析：



多模态融合优势显著：多模态融合模型的验证准确率（75.38%）显著高于仅文本（59.88%）和仅图像（64.88%）单模态模型，跨模态信息融合能有效提升情感分类性能。

单模态贡献存在差异：仅图像模态的准确率（64.88%）高于仅文本模态（59.88%），在该实验中，图像模态的单模态情感表达能力略强于文本模态，但两者均存在信息缺陷。

四、创新探索——多模态融合方法对比

在创新探索方向，选择开展多融合方法研究与对比。实验严格遵循“单一变量”原则，保持编码器、分类器、超参数（选择第一部分最优超参数）等完全一致，仅替换融合层，对比不同融合方法的性能差异。

1. 融合方法

Late Fusion（后期融合）：拼接融合、加权拼接融合、自注意力融合

特点：先独立提取文本、图像模态的高层特征，再在特征层完成融合操作；以特征拼接、可学习加权、自注意力特征交互为融合方式，无专门的显示跨模态语义对齐设计。

Early Fusion（早期融合）：特征投影融合

特点：在模态特征提取的早期阶段完成图文特征拼接，再对融合后的特征做统一的联合编码与投影处理，跨模态信息在特征学习初期即产生交互。

VLM 风格融合：CLIP 融合、BLIP 融合（基础/进阶版）

特点：基于视觉-语言模型（VLM）的核心设计，针对图文跨模态交互做专门优化；通过显式的跨模态特征对齐（CLIP）、视觉-语言引导式注意力交互（BLIP）挖掘模态间语义关联。

2. 实验流程

复用基线模型的训练与评估流程，批量训练所有模型，同时记录收敛轮数、损失趋势与验证集准确率。

3. 实验结果与分析

各模型最优验证集准确率对比

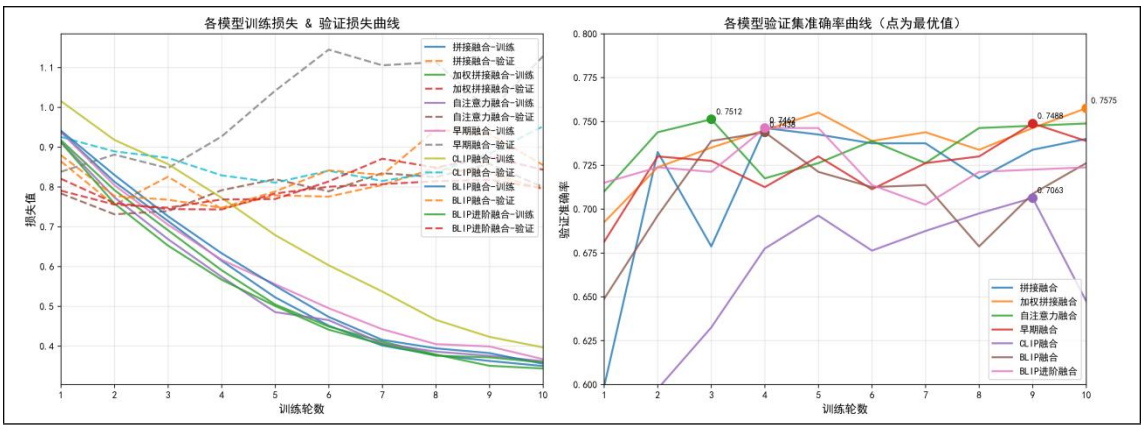
```
------------------------------------------------------------
拼接融合          | 最优准确率：0.7462（第4轮） | 对应训练损失：0.6147 | 对应验证损失：0.7448
加权拼接融合        | 最优准确率：0.7575（第10轮） | 对应训练损失：0.3438 | 对应验证损失：0.7967
自注意力融合        | 最优准确率：0.7512（第3轮） | 对应训练损失：0.6687 | 对应验证损失：0.7391
早期融合          | 最优准确率：0.7488（第9轮） | 对应训练损失：0.3992 | 对应验证损失：0.9907
CLIP融合        | 最优准确率：0.7063（第9轮） | 对应训练损失：0.4233 | 对应验证损失：0.8834
BLIP融合        | 最优准确率：0.7438（第4轮） | 对应训练损失：0.6333 | 对应验证损失：0.7472
BLIP进阶融合      | 最优准确率：0.7462（第4轮） | 对应训练损失：0.5670 | 对应验证损失：0.7683
```

整体性能对比：

最优：加权拼接融合（75.75%）、自注意力融合（75.12%），验证准确率均突破 75%，是实验中表现最好的融合方式；

良好：早期融合（74.88%），拼接融合（74.62%）、BLIP （进阶）融合（74.38%、74.62%），验证准确率在 74%~75% 区间，性能稳定；

较差：CLIP 融合（70.63%），验证准确率显著低于其他融合方式，泛化能力不足。



损失曲线分析：

训练损失：所有模型的训练损失均随训练轮数持续下降，说明模型能够正常收敛。且大部分融合方法收敛效果相差不大，其中仅 CLIP 训练损失下降最慢，

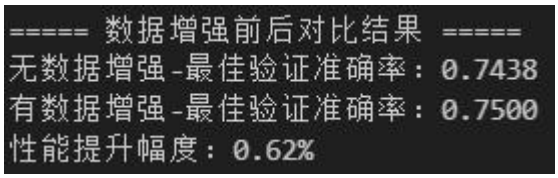
验证损失：不同融合方法的泛化能力分化明显，BLIP 融合和 Late Fusion 的训练与验证损失曲线贴合度高，泛化能力强，而早期融合的验证损失后期持续上升，泛化能力最差。

验证集准确率分析：

VLM 风格融合和 Late Fusion 的表现明显优于早期融合，其中加权拼接融合不仅准确率峰值最高，且后期稳定性最好，是本次实验中整体表现最优的融合方式。

4. 数据增强前后对比

使用上述得出的“加权拼接融合”方案进行实验。对训练集图像作数据增强处理：依次执行随机水平翻转（概率 0.5），随机旋转（±10° ）、色彩抖动（亮度/对比度/饱和度 ±0.2），通过随机化操作扩充训练样本多样性，并依照验证集准确率对数据增强前后做对比。



数据增强后模型验证准确率从 74.38% 提升至 75%，虽提升幅度 0.62% 但有效验证了增强策略的价值 —— 通过随机翻转、旋转、色彩抖动扩充了训练样本多样性，一定程度缓解了过拟合，让模型泛化能力略有提升。

五、补充部分

1. 代码实现时遇到的 bug

① 文件编码不匹配导致的解码失败。索引文件（train.txt、test_without_label.txt）的实际编码不是 utf-8，代码用 encoding="utf-8" 去读取，遇到了无法解码的字节。

索引文件实际是 ASCII 编码非 utf-8。

② data 数据集文本文件（.txt）存在混合编码，但代码中统一用 GBK 编码读取，导致不兼容的编码文件解码失败，触发 UnicodeDecodeError。

文本文件编码是混合场景—大部分是纯英文/数字/符号（ASCII 编码），少数含特殊字符（如俄语、日语等字符），是 IBM866 等编码。GBK 不兼容这些特殊编码的字节（例如文本中 4656.txt、4709 为日语）。

代码最终用 Latin-1 读取文本文件（具体编码可以参考 datasearch.ipynb）

2. 设计模型的原因与亮点

设计该模型是为贴合文本 - 图像双模态情感分类任务特性，针对基线拼接融合“等权处理模态、未考虑贡献差异”的局限，在保持编码器、超参等一致的单一变量原则下，通过轻量化优化（仅新增可学习权重）让模型自主适配图文模态的情感贡献度，同时为数据增强验证和多融合方法对比提供公平参照。

亮点是轻量化改造成本低，加权拼接较基线性能稳步提升且收敛稳定，可解释性强（符合图文模态贡献规律），能与数据增强协同增效，兼顾实验对比价值与工程实用性。

六、实验总结

本次实验聚焦文本 - 图像双模态情感三分类任务，分为基线构建与融合方法探索两部分：以 BERT+ResNet18 + 拼接融合构建基线，经超参调优（学习率 5e-5、权重衰减 3e-4）达 75.38% 准确率，消融实验证实双模态融合优于单模态；固定实验条件对比 7 种融合方法，加权拼接融合表现最优（75.75%），Late Fusion 类方法稳定性优于 Early Fusion 与部分 VLM 风格融合，CLIP 融合泛化不足。虽达成目标，但受数据集规模较小、VLM 融合适配不足等局限，未来可进一步优化。