实验五：多模态情感分类 10235501426 娄屹宏

Github 仓库地址：https://github.com/emperor239/AI_final

一、实验宏观设计

本实验一共 5 个模型：2 个基础版、2 个进阶版和 1 个创新版。

其中，2 个基础版模型的文本层均为 BERT-BASE、图像层均为 RESNET-18，但是在多模态融合上分别采用早期融合和晚期融合（因为晚期融合的模型独立性，加权和投票两种办法可以放在同一个模型文件内一起测试），从而分别构成了 2 个基础版模型。

2 个进阶版模型是分别通过微调 CLIP 和 BLIP 完成的，虽然验证集准确率较基础版模型提升 5%到 15%，但是训练过程出现明显的“准确率饱和问题”，进一步考察样本，发现了明显的“类别不平衡问题”（positive 和 neutral 的 F1 值偏差进一步增大，多数类主导了训练）。

因此，作者提出创新版模型，在 BLIP 的基础上添加了各种额外机制，来缓解“类别不平衡问题”。此外，作者还通过 Grad-CAM 热力图分析了各类别正判、误判的原因，并做出数据预处理和模型结构的创新与改进。训练显卡采用 NVIDIA RTX-5060。

二、通用模块

（一）文本清洗与图片增强

观察文本数据，发现 5 个典型结构：RT 转发符、@用户、#话题、网址、非英文与数字字符。需按序分步去除：过滤非 utf-8 的 error 字符、去除 RT、去除@用户、去除网址、去除#和非英文与数字字符。

观察图片数据，发现 3 个典型特征：长宽失衡、像素密度过高、主要内容居中。增强方法：将图片顶格放置在白色正方形中间，然后压缩成 224*224，并借助高斯模糊去噪，存储为模型需要的格式（ResNet18 需要 Tensor、CLIP 和 BLIP 需要 PIL），可以大幅降低训练时的磁盘 IO（通过任务管理器查看），以及图数据转换时所涉及的“CPU 加载-GPU 计算”与“内存-显存”交互，提高 GPU 利用率。

后续在观察了错误样本状况后，进一步改进了文本清洗与图片增强方案。在文本清洗方面，优化了含 apostrophe 的英语单词处理，防止把否定词切分。在图片增强方面，采用中心剪裁，仅保留主体信息，防止大量边缘留白，导致神经元失效。

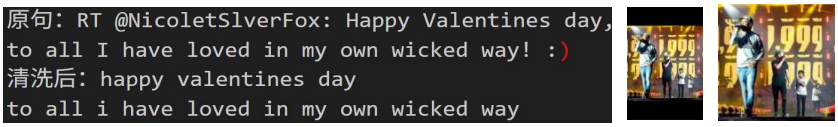

图 1 创新版数据读取的效果案例图（左为文本清洗，右为图像增强）

特别说明：spacy 库的停用词会破坏否定语义，词性还原不适配 BLIP 的预训练模型，因此不适用停用词和词性还原。

（二）通用优化方法和创新优化方法

通用优化方法	描述
Dropout正则化	让某一层的部分神经元失活（解冻层或表征层），防止过拟合（重点调参对象）
早停	当验证损失剧烈波动时停止，从而找到泛化能力最强的超参数
迁移训练	冻结大部分预训练层，只训练最后几层，从而适配本实验小数据集分类任务
归一化	按照图片RGB通道像素的均值和方差归一化为标准正态分布

创新优化方法	描述
标签平滑	将离散的硬标签变成连续的软标签，使得模型不把预测结果绝对化
样本对数平滑加权	样本类别比例约为6:1:3，极度不平衡，因此对样本采用对数平滑加权，让模型不被多数样本主导（重点调参对象）
学习率退火	当验证损失不再下降时，收缩学习率，帮助模型收敛到最优解
交叉验证	在训练集上用交叉验证训练模型，用验证集选择最优超参数，最后合并训练集和验证集，用最优超参数训练出最终模型，作用到测试集上，从而能用全所有已知数据
L2正则化	将图像表征和文本表征向量映射到同一个单位球内，防止较大表征向量的主导特征
neutral阈值	对于无法可靠地确定为positive或negative的样本，分类为neutral

表 1 通用优化方法（左，所有模型均采用）和创新优化方法（右，仅在创新版模型采用）

类别不平衡问题。各个模型的第一轮训练的验证损失总是 59.75%，而本实验的 positive、neutral、negative 三个类别约为 6：1：3，positive 比例正好匹配，这是严重的类别失衡问题，模型容易受多数类主导，所以训练时需要额外对不同类别的样本加权（参考机器学习 AdaBoost 方法的思想）。此外，数据要按标签分层抽样，保证训练/验证集标签分布一致。

过拟合问题。由于不是物品分类而是情感分类，会出现严重的训练集和验证集过拟合，

解决方法是 Dropout+早停+交叉验证，即观察前 5 轮，训练损失和验证损失应接近，且一起中等幅度下降，并在后期小幅度波动。创新方案额外采用交叉验证和带权重的损失（见图）。

三、模型结构的简要介绍

（一）四个基础模型的结构概要

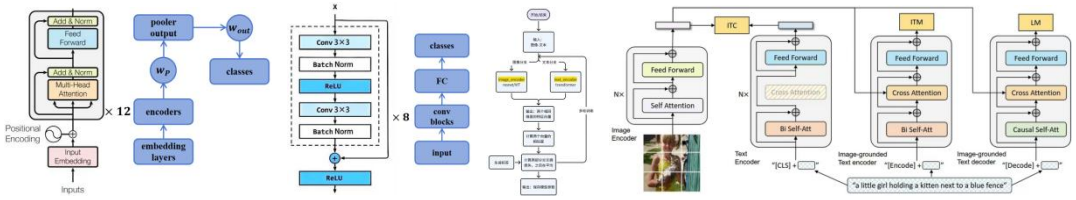

图 2 BERT、RESNET-18、CLIP、BLIP 结构示意图（参考文献见 README.md）

BERT 是多层双向 Transformer Encoder，含多头注意力、前馈网络，配嵌入层和残差归一化；RESNET-18 用 4 个残差块（3 × 3 卷积、残差连接），尾接全局池化和全连接；CLIP 在图像侧用 RESNET 在文本侧用 Transformer Encoder，在图像侧用 RESNET，无跨模态交互；BLIP 在文本侧用 Transformer，在图像侧用 ViT，添加跨模态注意力。

（二）融合逻辑

早期融合则直接将两个模型的 embedding 融合，共享一个分类器。

晚期融合分为加权平均和投票机制。加权平均，即给文本和图像的概率分配权重，最终概率最高的作为预测标签；投票机制，即文本和图像各自预测标签，投票取概率更高的模态结果。

但加权平均前提是两个特征的分布一致，否则会出现一方特征向量被另一方特征向量淹没的问题。因此需要 L2 归一化，将每个样本的特征向量归一化到单位球面上，使得图文特征的尺度统一，保证两者在融合时尺度与空间位置平等，不会出现一方被压制的情况。

（三）本实验的模型

模型名称	使用到的模块
基础版+早期融合	上文所述的通用优化方法、BERT、RESNET-18、早期融合方法
基础版+晚期融合	上文所述的通用优化方法、BERT、RESNET-18、两种晚期融合方法
进阶版+CLIP	上文所述的通用优化方法、CLIP
进阶版+BLIP	上文所述的通用优化方法、BLIP
创新版+BLIP	上文所述的通用优化方法、BLIP、标签平滑、样本对数平滑加权、交叉验证、学习率退火、交叉验证、L2正则化、neutral阈值

表 2 本实验的模型中使用到的优化模块（具体模块受限于篇幅请对应表一查看）

因为本实验的图文数据量过于稀少，所以 5 个模型均采用了迁移训练的方式。作者冻结预训练模型的前大多数层，但是仅将最后两层解冻，进行模型微调。但由于各个预训练模型每一层的参数量都很大，为了避免过拟合，作者对各个迁移训练模型的架构进行了修改，即给每个解冻层额外加一个 Dropout 层，且在融合之前，对图文特征分别做 L2 归一化，最后才用一个分类器把所有结果映射为情感类别，从而缓解图文情感分类中的过拟合问题。

四、实验数据测定与分析

模型名称	总参数量	训练参数量	解冻层数	epoch	positive F1	neutral F1	negative F1	验证集准确率	Macro F1	Weighted F1	训练时长
基础版+早期融合	121515331	97668675	图像2层+文本2层	6	0.7992	0.2321	0.6509	0.7125	**0.5607**	0.6955	245.33秒
以上模型消融实验：仅文本保留	121515331	97668675	文本2层	7	0.7201	0.0000	0.4484	0.6050	0.3895	0.5636	267.95秒
以上模型消融实验：仅图像保留	121515331	97668675	图像2层	15	0.7522	**0.2797**	0.5088	0.6412	0.5136	0.6302	494.42秒
基础版+晚期融合(加权平均)	121501486	83479140	图像2层+文本2层	略	0.5183	0.2157	**0.7944**	0.6963	0.5095	0.6515	略
基础版+晚期融合(投票机制)	121501486	83479140	图像2层+文本2层	略	0.5296	0.2500	**0.7993**	0.7037	0.5263	0.6614	略
以上模型消融实验：仅文本保留	109679875	71666947	文本2层	6	0.6028	0.2545	**0.7891**	0.7025	0.5488	0.6775	90.94秒
以上模型消融实验：仅图像保留	11821611	11812139	图像2层	6	0.3784	0.2314	0.7714	0.6488	0.4604	0.5978	47.66秒
进阶版+CLIP（add融合）	151409412	21267971	图像2层+文本2层	6	**0.8354**	**0.2804**	0.6968	**0.7600**	**0.6042**	**0.7359**	212.58秒
进阶版+BLIP（add融合）	224792068	33538563	图像2层+文本2层	12	**0.8380**	0.0235	0.6901	**0.7500**	0.5172	**0.7085**	1488.09秒
创新版+BLIP（统计第一折）	224858116	33604611	图像2层+文本2层	16	**0.8069**	**0.4138**	0.6539	**0.7262**	**0.6249**	**0.7204**	约150分钟

表 3 本实验 5 个模型+4 个消融实验的实验数据（各指标排名前三的高亮并加粗）

可以看到，本次实验将解冻层数为 2 作为最核心的固定变量，从而控制变量进行实验。其中特别保证了“进阶版 BLIP”与“创新版 BLIP”的总参数量和训练参数量几乎一致，从而可以看到我们创新模块的效果。

此外，由于数据集上有严重的类别不均衡问题，要使用 F1-score 才能科学地评判模型的效果，且整体评价要看 Weighted F1，而不是 Macro F1，因为类别不均衡。

由于实验控制变量，前几个模型无法达到最优效果，会出现病态的训练曲线，请谅解。

（一）结论一：双模态融合效果一般会优于单模态

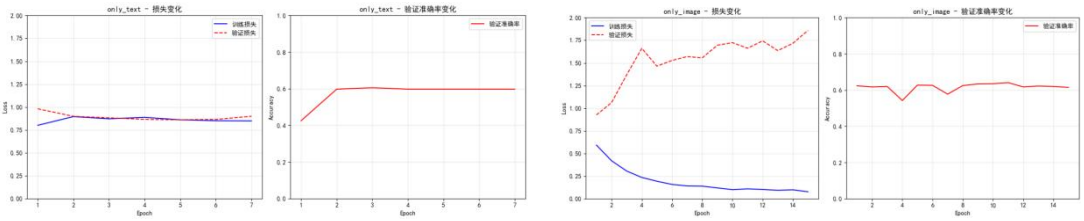

图 3 “基础版+早期融合”仅文本的训练曲线（左一、二）和仅图片的训练曲线（右一、二）

比较完整的双模态模型和仅文本/仅图像的消融模型（表 3 的前 3 行）：“基础版+早期融合”的 Weighted F1 为 0.6955，而仅文本版本骤降至 0.5636，仅图像版本也只有 0.6302。从其训练曲线也可以看出，消融模型的单模态存在严重的过拟合问题和准确率平台期问题，而且明显被不平衡的类别干扰了。不过，文本与图像的双模态融合能一定程度上提升模型的分类性能和鲁棒性。消融实验说明多模态缺一不可的重要性。

（二）结论二：文本和图像模态在不同模型架构中的相对重要性不同

晚期融合由于训练独立，最终结果不会相互干扰，因此融合模型与各个模态的单分支模型的 Weighted F1 误差在可接受范围内，所以各个模态对融合模型的影响也几乎相互独立。

（三）结论三：类别不平衡问题导致的多数类主导训练的问题干扰严重

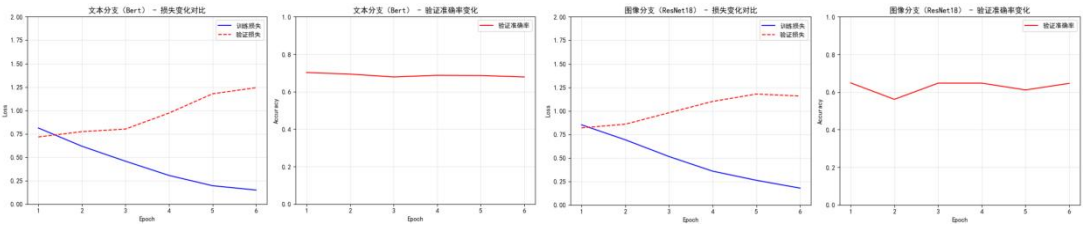

图 4 “基础版+晚期融合”仅文本的训练曲线（左一、二）和仅图片的训练曲线（右一、二）

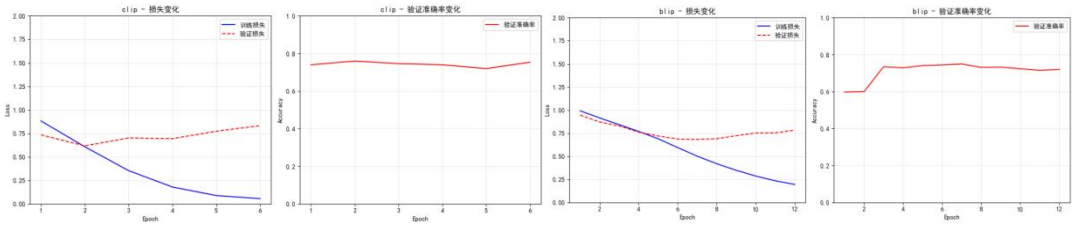

图 5 CLIP 的训练曲线（左一、二）和 BLIP 的训练曲线（右一、二）

可以看到，以上的这些模型都产生了明显的过拟合问题与准确率平台期问题，并且在 neutral F1 上的值低于 33%，训练被多数样本主导，必须通过引入依据样本类别权重而计算的损失来引导训练。

尤其是中性情感。在所有实验中，neutral F1 普遍偏低，甚至出现 0。这表明中性情感的识别是该任务中的主要瓶颈，其原因可能在于：中性情感的表达更隐晦，缺乏强烈的文本或视觉特征；数据集中中性样本的数量可能较少或标注质量不高，导致模型学习不足。

（四）结论四：早期融合在本实验中比晚期融合好

“基础版+早期融合”的验证集准确率为 0.7125、Weighted F1 为 0.6955，而“基础版+晚期融合（加权平均）”的验证集准确率仅有 0.6963、Weighted F1 为 0.6515。与此同时，早期融合的 positive F1 也远高于晚期融合，说明其情感捕捉能力更强，即早期融合能更充分地利用两种模态的原始特征，比简单的晚期拼接或相加更有效。

（五）结论五：下游任务主导预训练模型的参数量控制

“进阶版 CLIP”在各项关键指标上全面优于进阶版 BLIP，尤其在 neutral F1 上，CLIP（0.2804）远高于 BLIP（0.0235），显示出对最难分类的细粒度中性情感更强的识别能力。但我一开始查阅到的内容明明是 BLIP 比 CLIP 好，可能是控制了变量的原因。

“进阶版 CLIP”（总参数量 1.5 亿）在各项指标上全面超越基础版模型（总参数量 1.2 亿）。此外，进阶版 BLIP 模型（总参数量 2.24 亿）的表现却不如参数量更小的 CLIP，这说明模型性能的提升不仅仅取决于参数量的大小，更关键的是预训练模型与下游任务的适配性。

（六）结论六：测试集需要额外的 neutral 判定机制

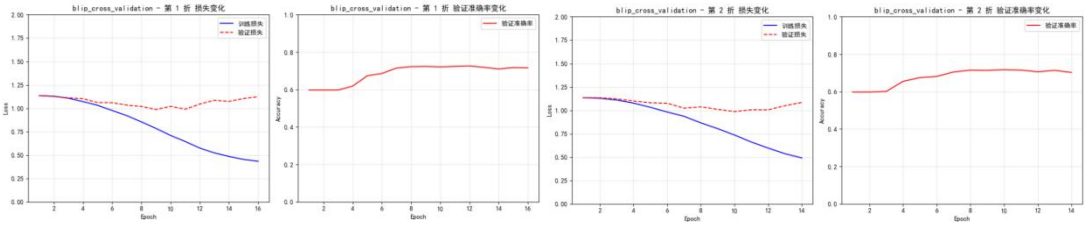

图 6 创新版+BLIP 的训练曲线（左一、二为第一折，右一、二为第二折，其它折类似）

由于实在难以高效解决 neutral 样本失衡的问题，通过一系列机制的优化，也只是得到了较为健康的训练曲线，但依旧存在准确率停滞和泛化能力不足的过拟合状况。作者认为，传统图片分类在图片中往往体现相同的视觉特征，比如都有轮子和翅膀的形状，传统文本分类一般是有数百字篇幅的新闻报道。而情感分类则与图片、文本数据不是直接相关联的，即因此难度很大，作者只有能力做到这样的损失曲线效果。

为了提高测试集的效果，在测试集输出各个分类的概率后，仍会对 positive 和 negative 两个类别的概率进行阈值判定，一旦两者的概率低于某个阈值，则直接分类为 neutral，从而保证 neutral 的预测可靠性。“创新版+BLIP”模型就是为了缓解类别不均衡导致的问题，优势就在于缓解 neutral 的漏判问题，比其他模型的 neutral F1 都几乎高出 1 倍。

五、基于 Grad-CAM 热力图的样本可视化与改进

本节可视化为额外的插入内容，用以说明模型判断 positive 的可能的思路，也用以说明类别不平衡问题的本质是经验风险最小化，从而将“创新版+BLIP”改进为了对数平滑加权。

（一）Grad-CAM 的原理与解释性

Grad-CAM 是针对卷积神经网络的可视化技术，核心是通过目标类别对卷积层输出特征图的梯度，计算每个特征图的权重，再加权求和得到类激活热力图。计算步骤大致为：前向传播得到目标类别预测值；反向传播计算该类别对最后一层卷积特征图的梯度；对梯度在空间维度求平均，得到每个特征图的重要性权重；权重与对应特征图加权求和，经 ReLU 激活（保留正贡献），上采样至原图尺寸，生成热力图。以下以 positive 的正判和误判问题为例。

（二）正例

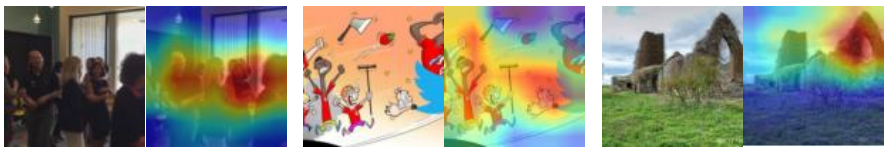

图 7 真实为 positive 且被模型猜测为 positive 的三对典型样本（每对左原图，右热力图）

作者发现了三种会被判定为 positive 的情况：群像、夸张动作、风景实物。

第一种就是如左侧这一类包含多个人群像的图片。显然，模型也将注意力锁定到了各个人物的脸和肩膀上。这样的图片在现实层面上往往预示着聚会活动，因此往往是 positive 的。模型经验性地对包含多目标的图片判定为 positive，一般不会出错。

第二种就是如中间这一类包含夸张动作的图片。此时，模型将注意力锁定到了横跨多个卷积块的物体之上，并将动作的受体锁定到了两个人与动物上。这样的图片在现实层面上往往意味着活跃、热闹和开朗的氛围，多目标之间的互动性强，因此有理由判定为 positive。

但第三种就是如右侧这种聚焦风景实物的图片，如在本例中就聚焦在岩石建筑上，这种借景抒情、借物抒情的图片其实并不好揣测情感，但由于 positive 占主要类别，因此，模型宁愿判定为 positive，这是经验风险最小化的决策过程，后续反例部分会分析该原理。

（三）反例

作者还准备了一些反例的对比图，来证明在有明显焦点的图片中，模型其实并没有真正学明白人物表情，只是在进行经验风险最小化的猜测。

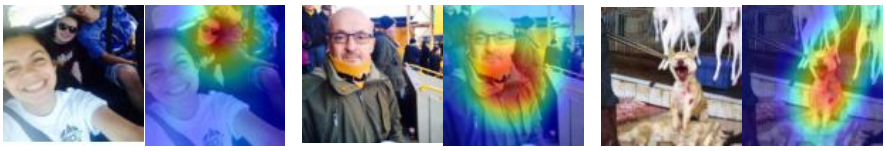

图 8 从左到右依次为 positive、neutral、negative，但模型都猜测为 positive

在左侧的正例中，模型锁定了人脸，因此判定为为 positive。但是在中间的 neutral 中，这个人并没有微笑，反而面部神情严肃，然而模型依旧锁定其人脸，然后直接判定为 positive。由此可以断定，模型仅能识别物体，但还无法细致地学习嘴巴是否真的是上扬的表情特征。在右侧的 negative 案例中，猫的神色狰狞，但是模型只要聚焦到面部，就简单地识别为 positive，这是一种经验风险最小化策略，类似于 KNN 中 K 设置较大时的场景，即模型的考量简单，只要保证将领域内最多的类别视为本领域的类别，即可保证本领域的经验误判率最小化。

因此，为了避免模型进行经验风险最小化的猜测，作者在“创新版+BLIP”内着重添加了标签平滑损失和对数平滑加权，从而迫使模型进一步学习更细致的图像特征。

六、实验过程中各种 bug 与对应的解决方式

网络防火墙 bug。调参时会因连续爬取 huggingface 网站的预训练模型而被墙。可以切换网络代理，并将预训练模型下载至本地缓存，使模型读取本地缓存。

GPU 与 CPU 的兼容。DataLoader 设置$num_workers = 2$，可以设置$pin_memory = True$，将数据加载到 CPU 的锁定内存中，后续迁移到 GPU 时减少内存拷贝开销；同时设置$prefetch_factor = 2$，提前加载下一批数据，提高内存复用率，进一步减少 GPU 等待。以上在$num_workers = 1$ 时会出 bug，所以只能二选一。

批处理与多进程兼容 bug。多模态模型的输入是图文双特征。本来想自定义 DataLoader 的批处理函数 collate_fn，但这会导致 DataLoader 无法多进程加载数据，使得训练时 GPU 利用率仅 35%至 55%。可以在 Dataset 的__getitem__中返回字典，即可让 PyTorch 默认的 collate_fn 对字典中每个 value 进行堆叠，从而适配图文双特征输入，同时兼容多进程加载数据，保证 GPU 满载（实测 93%到 100%）。

迁移训练的层解冻 bug。迁移训练时可能出现无法解冻的问题，其实是没把文本处理器和图像处理器分开解冻。可以分别独立调控图像和文本的解冻层数。

张量维度 bug。大部分 bug 是一些张量维度的兼容问题，较为具体，不赘述。

七、总结

本次多模态情感分类实验构建了 5 个核心模型（2 基础+2 进阶+1 创新），覆盖了不同的融合策略，并采用控制变量法进行模型之间的横向对比。针对文本图像数据的问题进行了清洗，针对类别失衡的问题采用对数平滑加权、5 折交叉验证和分层抽样进行一定程度的缓解，还结合了 Dropout、早停、标签平滑等优化策略进一步优化模型。实验证实双模态融合一般优于单模态，预训练模型下游任务适配性比参数量更关键（CLIP 优于 BLIP），中性情感识别（F1 普遍低于 33%）是核心瓶颈。同时通过 Grad-CAM 可视化，对模型进行进一步改进，解决实际问题，提升实验结果的实用性，为同类多模态情感分类任务提供可复用经验。

但本次实验也有一个技术难点，就是小样本下“类别不平衡问题”的处理，若进一步实验，作者会考虑对多数类进行欠采样，但样本量本来就很少，没怎么敢在本次实验中尝试。

就情感分类而言，有直接抒情（往往有明显的人像笑脸和文本）、借景抒情、借物抒情。但后两者较委婉，小规模小样本的神经网络对这一类的图文数据会摸不着头脑。