# 实验五：多模态情感分析

姓名：袁凡　　学号：10214304412

GitHub 地址: https://github.com/Leofyfan/multimodal_sentiment_analysis.git

## 1　实验说明

多模态融合模型实现情感分析的范式为：特征提取 + 特征融合 + 情感分类。特征提取可分为两类：使用单模态预训练模型分别提取对应模态特征，多模态预训练模型提取特征（如 CLIP 等）。特征融合过程可采用简单的拼接、相加等。然而，简单的融合方式无法关注不同模态特征的差异性，甚至会破坏各模态的语义信息。因此在融合过程中需要额外关注如何缩小不同语义子空间中的分布差距，同时保持模态特定语义的完整性。

**本实验基于单模态预训练模型（Bert-Base、Swin-Transformer-V2-Base）分别提取文本、图像特征，并采用 6 种不同的融合方式（拼接、相加、注意力、拼接 + 注意力、相加 + 注意力、transformer-encoder）进行特征融合**，最后使用全连接层进行情感分类。

此外，本实验的数据集存在严重的类别不平衡问题（Negative: 29.83%, Positive: 59.70%, Neutral: 10.47%）。在模型评估时发现，模型在 Positive 和 Negative 类别的预测表现远优于 Neutral 类别。基于此，本实验对数据进行了数据增强：对文本进行同义词替换、随机插入删除，对图像进行对比度、亮度调整和旋转等增强操作；**并提出了自适应类别平衡损失函数（adaptive_class_balanced_loss）**，以改善模型在不平衡类别预测的表现。

## 2　模型搭建

### 2.1　模型训练流程

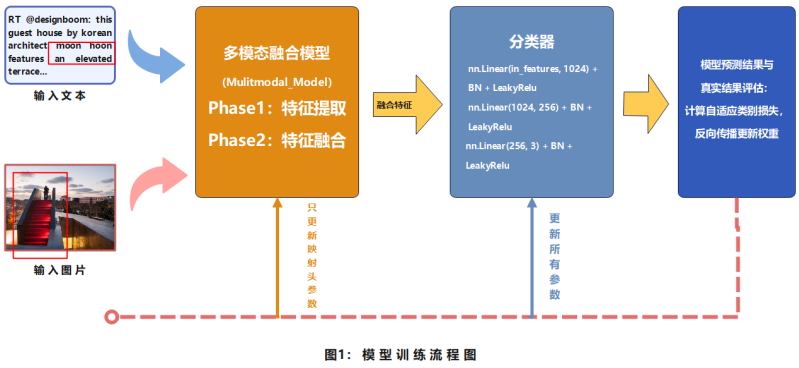


图 1: 模型训练流程图

**流程说明：**将文本和图像对输入多模态融合模型中，经过特征提取阶段 (phase1) 和特征融合阶段 (phase2) 得到融合的特征向量，该特征向量经过分类器得到模型预测分类结果。最后与真实结果评估，计算自适应类别平衡损失，并进行反向传播更新模型参数。

**注意：**这里反向传播过程中会更新分类器的所有参数，而多模态融合模型只会更新部分参数（文本和图像的映射层参数），预训练模型（Bert-Base、Swin-Transformer-V2-Base）的主体参数会被冻结不进行更新。如果融合方式采用了注意力和 encoder 的话，也会对注意力头和 encoder 进行参数权重更新。这样做可以降低模型训练成本、避免过拟合、保留预训练知识，在训练数据量较小的场景下，也能快速适配多模态情感任务。

## 2.2 特征提取和特征融合流程

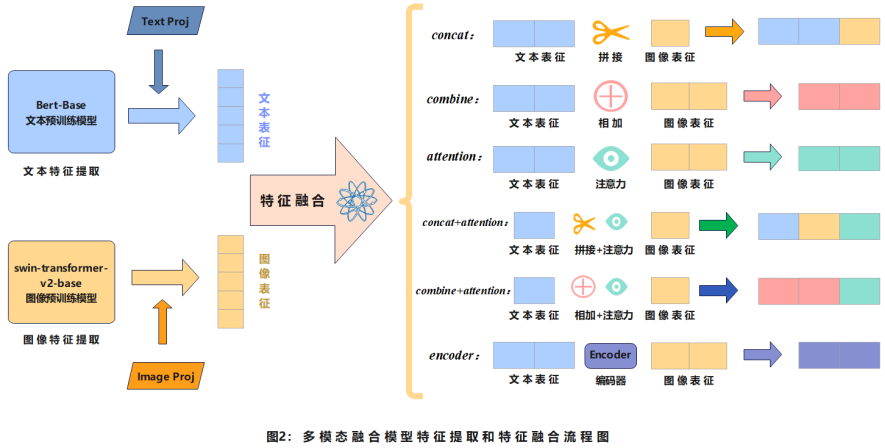


图2：多模态融合模型特征提取和特征融合流程图

图 2: 特征提取和融合流程图

**流程说明：**利用 Bert-Base 和 Swin-Transformer-V2-Base 分别提取文本和图像特征，并将该特征经过文本映射层（Text Proj）和图像的映射层（Image Proj）得到新的特征向量。在进行特征融合时，可根据训练配置采用拼接、相加、注意力、拼接 + 注意力、相加 + 注意力、transformer-encoder 六种融合方式得到对应的融合特征。

**注意：**为了关注不同模态特征的差异性，concat 融合方式对应的文本和图像表征的维度是不同的。本实验认为越重要的模态对应的表征维度越大，在实验过程中发现文本模态的重要性是大于图像模态的。在超参数搜索和模型训练日志可以发现上述规律（后面的消融实验也可以印证这一点），过程详见 search 文件夹和 logs 文件夹。

## 2.3 自适应类别平衡损失函数

本实验中采用自适应平衡损失函数（adaptive_class_balanced_loss），该损失函数通过动态调整类别权重和焦点参数，同时结合边界增强项，实现了对难样本的自适应学习和类别边界的优化。自适应类别平衡损失函数的数学表达式如下：

$$\mathcal{L}_{acb} = \alpha \cdot \mathcal{L}_{focal} + \beta \cdot \mathcal{L}_{boundary} = \alpha \cdot \frac{1}{N} \sum_{i=1}^{N} w_c (1 - p_t)^{\gamma_c} (-\log(p_t)) + \beta \cdot e^{-2(p_1 - p_2)}$$

**参数说明：**

- $\mathcal{L}_{focal}$ 是自适应焦点损失函数，$p_t$ 是正确类别的预测概率，$w_c = 1 + 5\log(1 + \frac{\sum_{i \in c}(1-p_t^i)}{N_c})$ 是类别 $c$ 的自适应权重，$\gamma_c = 1.5 + 3\log(1 + \frac{\sum_{i \in c}(1-p_t^i)}{N_c})$ 是类别 $c$ 的动态焦点参数。
- $\mathcal{L}_{boundary}$ 是边界损失（增强项），$p_1, p_2$ 分别是最高和次高预测概率
- $\alpha$ 是自适应焦点损失的权重系数，$\beta$ 是边界损失的权重系数。

# 3 实验结果

注：模型超参数搜索算法实现见 search 文件夹、训练日志见 logs 文件夹。模型训练损失和指标记录见 Wandb 项目页面。限于篇幅，实验报告只展示部分重要实验结果。

## 3.1 融合方式效果对比

表 1: 不同融合方式的性能对比

| 融合方式 | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Concat | 0.7188 | 0.6185 | 0.6049 | 0.6066 |
| Combine | 0.7275 | 0.6441 | 0.5820 | 0.5939 |
| Attention | 0.6450 | 0.5512 | 0.5409 | 0.5455 |
| Attention+Concat | **0.7550** | **0.6938** | 0.6512 | **0.6713** |
| Attention+Combine | 0.7538 | 0.6889 | **0.6526** | 0.6703 |
| Transformer-Encoder | 0.7525 | 0.6912 | 0.6498 | 0.6698 |

**实验结果说明**：

- Attention+Concat 融合方式在 Accuracy、Precision、F1-score 上取得了最好的效果，这表明注意力机制和特征拼接的组合能够更好地捕捉模态间的互补信息。transformer-encoder 的表现接近 Attention+Concat，说明利用编码器也能有效地建模模态间的交互关系。
- 单独使用 Attention 的效果最差，说明仅依靠注意力机制会丢失一些重要的模态特定信息。加入注意力的融合方式（Attention+Concat、Attention+Combine），模型性能普遍优于简单的融合方式（Concat、Combine）。说明注意力机制可以捕捉到不同模态信息的关联信息，该关联信息可以作为额外信息辅助模型预测，从而提升模型性能。
- Combine 融合方式的性能不是特别差，说明该文本 (Bert-Base) 和图像 (Swin-Transformer-V2-Base) 的预训练模型对文本和图像的表征具有较高的对齐度或者说该预训练模型的通用性较强。因为直接相加的方式会破坏不同模态的语义信息，从而导致性能下降。

## 3.2 损失函数效果对比

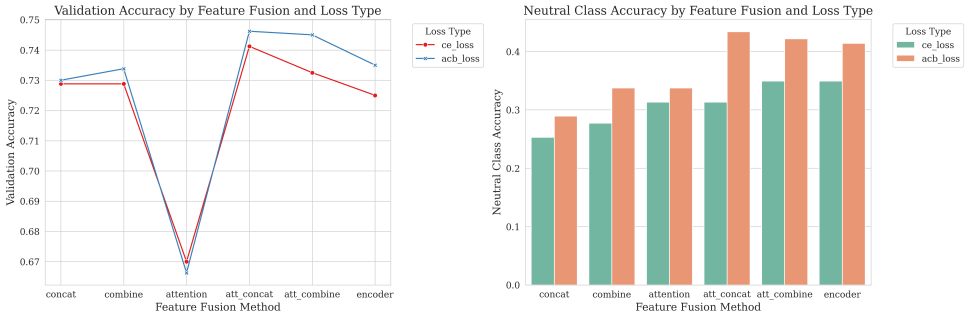


图 3: 损失函数对比

**实验结果说明：**

- 在验证集的整体类别准确率上，自适应类别平衡损失函数（acb_loss）相比传统的交叉熵损失函数（ce_loss）在多数特征融合方式都展现出了更优的性能：如在 attention + combine 方式下从 73.25% 提升至 74.50%，在 encoder 方式下从 72.50% 提升至 73.50%。尤其对于不平衡的 Neutral 类，acb_loss 的提升更为显著，特别是在 attention + concat 融合方式下，准确率从 31.33% 大幅提升至 43.37%。
- acb_loss 的在提升整体准确率的同时，显著改善了占比最小（10.47%）的 Neutral 类的预测效果，这说明该损失函数成功解决了类别不平衡问题；其次，通过边界增强项的设计，acb_loss 增强了模型对不同类别的区分能力，使得模型在各个融合方式下都取得了更加平衡和优秀的性能表现，展现出了良好的泛化能力；也说明 acb_loss 的损失函数设计是合理的。

# 4 消融实验

表 2: 不同融合方式的模态消融实验结果

| 融合方式 | 模态 | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| Concat | Text | 0.7025 | 0.6205 | 0.5742 | 0.5877 |
| | Image | 0.6663 | 0.5796 | 0.5310 | 0.5461 |
| | Text+Image | **0.7188** | **0.6185** | **0.6049** | **0.6066** |
| Combine | Text | 0.6987 | 0.6319 | 0.5338 | 0.5309 |
| | Image | 0.6600 | 0.5743 | 0.4995 | 0.5152 |
| | Text+Image | **0.7275** | **0.6441** | **0.5820** | **0.5939** |
| Attention | Text | **0.7025** | **0.6148** | **0.5977** | **0.6038** |
| | Image | 0.6538 | 0.5806 | 0.5302 | 0.5456 |
| | Text+Image | 0.6450 | 0.5512 | 0.5409 | 0.5455 |
| Attention+Concat | Text | 0.7050 | 0.6399 | 0.5884 | 0.5942 |
| | Image | 0.6650 | 0.5911 | 0.5243 | 0.5411 |
| | Text+Image | **0.7550** | **0.6938** | **0.6512** | **0.6713** |
| Attention+Combine | Text | 0.7087 | 0.6425 | 0.5927 | 0.6168 |
| | Image | 0.6650 | 0.5921 | 0.5227 | 0.5553 |
| | Text+Image | **0.7538** | **0.6889** | **0.6526** | **0.6703** |
| Transformer-Encoder | Text | 0.7087 | 0.6417 | 0.5935 | 0.6167 |
| | Image | 0.6412 | 0.5552 | 0.5227 | 0.5385 |
| | Text+Image | **0.7525** | **0.6912** | **0.6498** | **0.6698** |

**消融实验结果说明:**

- **单模态与多模态的差异**：在大多数融合方式下（attention 除外），无论是文本（Text）还是图像（Image）单模态，其分类性能均低于多模态融合（Text+Image）。尤其是在 Attention+Concat 和 Attention+Combine 方式中，多模态融合的 Accuracy 分别达到 0.7550 和 0.7538，显著高于单模态。这表明单一模态的信息有限，难以全面捕捉情感特征；而多模态融合能够有效结合文本和图像的互补信息，提升分类效果。
- **模态的重要性**：文本模态在大多数情况下表现优于图像模态，例如在 Concat 方式中，文本单模态的 Accuracy 为 0.7025，而图像单模态为 0.6663。这表明在该情感分类任务中，文本信息更具判别性，含有更丰富的语义信息。但结合图像模态后性能进一步提升，说明图像模态提供了额外的有用信息。

# 5 实验总结与改进

**总结： 模态语义对齐问题：**实验提出了六种融合方式，利用注意力机制和特征拼接结合的方式在缩小不同语义子空间中的分布差距同时，也保证了模态特定语义的完整性；有效地捕捉了模态间的互补信息。**数据集存在严重的类别不平衡问题：**实验提出了自适应类别平衡损失函数（adaptive_class_balanced_loss），有效缓解了类别不平衡问题，显著提升了模型在不平衡类别上的预测性能。

**改进：**本实验采用的范式为单模态预训练模型分别提取再进行特征融合，缺乏对多模态联合语义的理解和建模能力。而多模态预训练模型（如 CLIP）通过在大规模多模态数据上进行联合训练，能够更好地捕捉文本和图像之间的语义对齐关系，从而在特征提取和融合过程中表现出更强的泛化能力和鲁棒性。未来的实验可以尝试使用 CLIP 等多模态预训练模型替代单模态预训练模型，直接提取联合的多模态特征。这样可以避免特征融合过程中的信息损失，并充分利用多模态联合语义的优势。

# 参考文献


[1] LIU Qiwei, LI Jun, GU Beibei, ZHAO Zefang. TSAIE: Text Sentiment Analysis Model Based on Image Enhancement[J]. *Frontiers of Data and Computing*, 2022, 4(3): 131-140. CSTR: 32002.14.jfdc.CN10-1649/TP.2022.03.010. DOI: 10.11871/jfdc.issn.2096-742X.2022.03.010

[2] DEVLIN Jacob, CHANG Ming-Wei, LEE Kenton, TOUTANOVA Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. *arXiv preprint arXiv:1810.04805*, 2018. DOI: 10.48550/arXiv.1810.04805

[3] LIU Ze, HU Han, LIN Yutong, et al. Swin Transformer V2: Scaling Up Capacity and Resolution[C]. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022: 12009-12019. DOI: 10.1109/CVPR52688.2022.01171