当代人工智能实验五实验报告

本项目已推送 github 仓库 https://github.com/SJF-ECNU/AiLabFinal.git

一.实验任务

给定配对的文本和图像，预测对应的情感标签。

三分类任务：positive, neutral, negative。

设计一个多模态融合模型，自行从训练集中划分验证集，调整超参数，预测测试集（test_without_label.txt）上的情感标签。

二.数据分析

我们首先对数据集进行一个简单的分析，这里我编写了代码 data_analysis.py 进行分析，其中最我最关系的部分如下

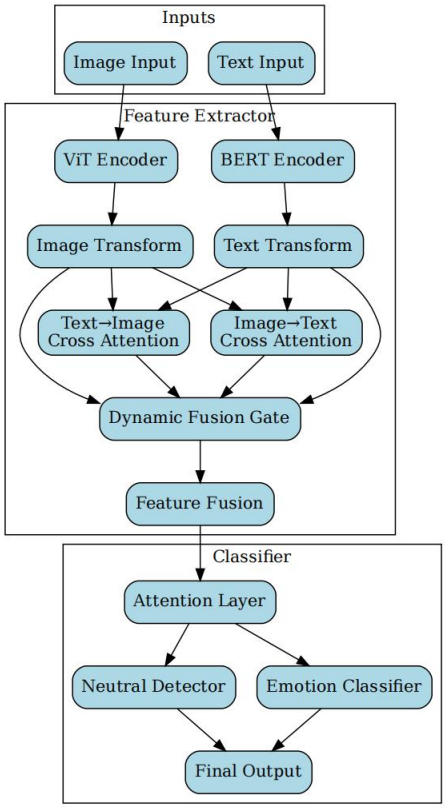
```
=== 标签分布 ===
positive: 2388 (59.70%)
negative: 1193 (29.83%)
neutral: 419 (10.47%)
```

我们可以发现一个很明显的情况，数据集是偏态的，其中，positive 类别的样本数量明显多于 negative 和 neutral 类别的样本数量。这种不平衡的数据分布可能会对模型的训练产生影响，使得模型更容易偏向于预测多数类别，即 positive 类别。为了处理这个问题，我可能需要采取一些措施，比如使用数据增强技术来增加少数类别的样本数量，或者在训练过程中使用权重调整策略来平衡不同类别的贡献。

三.模型的构建

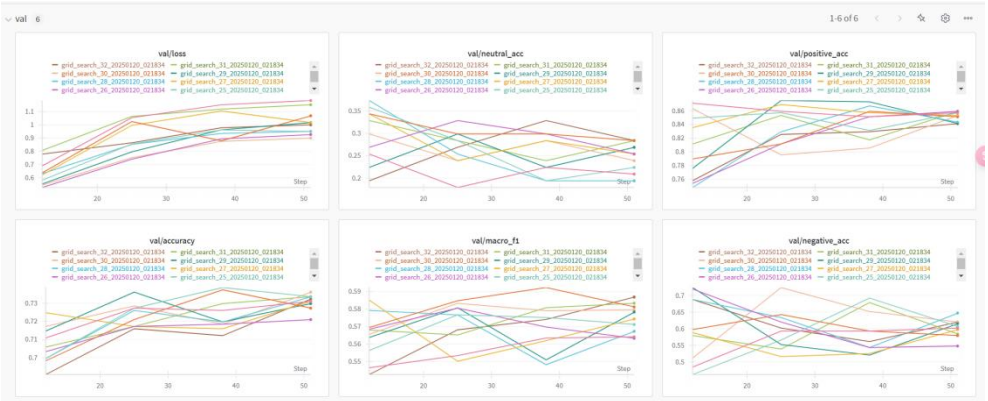
为了获取两种数据的特征，我需要两种编码器，文本编码器使用预训练的 BERT 模型提取文本特征，图像编码器使用预训练的 ViT(Vision Transformer)模型提取图像特征，然后进行交叉注意力融合，获得四种特征，包括原始文本特征、原始图像特征、文本注意力特征和图像注意力特征，并通进行动态融合，融合为最终的权重。得到权重后，为了应对数据集的不平衡问题，我将分类器分为两个部分，一个单独的 Neutral Decoder 专门用于判断是否为中性情感，最后使用 Sigmoid 函数输出 0-1 之间的概率值。同时 Emotion Classifier 用于用于区分积极/消极情感，并在最后进行组合输出。

模型结构图如下



四.超参数搜索

首先，我对于数据集进行清洗，并对数据集进行 8:2 的比例进行训练集和验证集的划分，划分后我进行了超参数的搜索，搜索的部分结果如下



由此我得到了最佳的超参数组合，并保存在了 config.py。

五.模型训练过程的问题和思考

接下来所有的内容中，用的模型是上一步找到的最佳配置训练得到的模型，为红色折线

1.数据集增强

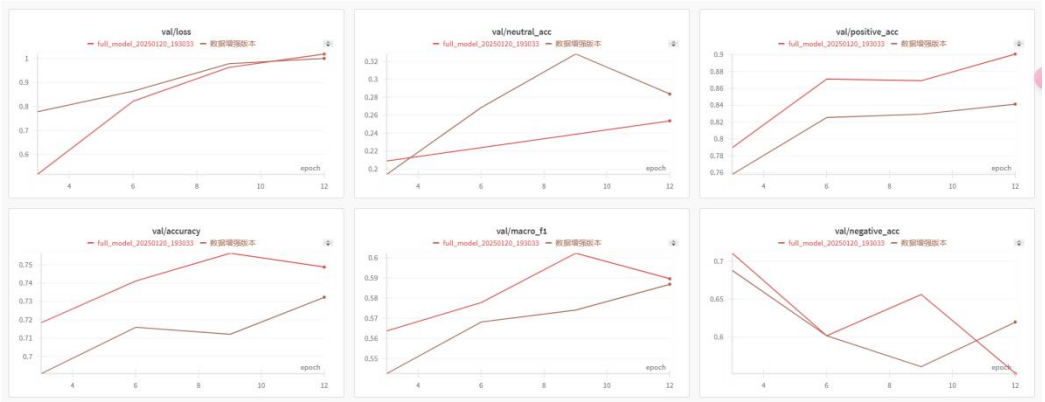
我在训练的过程中查看过不同 label 对应的准确率，这里展示一个例子

```
Epoch 9:
Average training loss: 0.0395

Training Metrics:
Overall Accuracy: 0.9655
Macro F1: 0.9655

Class-wise Accuracies:
Negative: 0.9594
Neutral: 0.9830
Positive: 0.9543

Validation Metrics:
Overall Accuracy: 0.6932
Macro F1: 0.5327

Class-wise Accuracies:
Negative: 0.6244
Neutral: 0.1791
Positive: 0.7917
```

事实上我们可以发现，neutral 的 val 准确率确实因为数据集信息量太小而非常低，因此我尝试过进行数据集的增强，如图像的旋转、明暗，文本的同义词替换，经过这样操作后，neutral 类的准确率有所上升，但是整体的准确率降低了，不过为了同时我计算了 Macro F1，用于评估平等对待各类别的情况下的性能，公式如下

- Precision = TP / (TP + FP) 真正例/（真正例+假正例）
- Recall = TP / (TP + FN) 真正例/（真正例+假负例）
- F1 = 2 *(Precision* Recall) / (Precision + Recall)

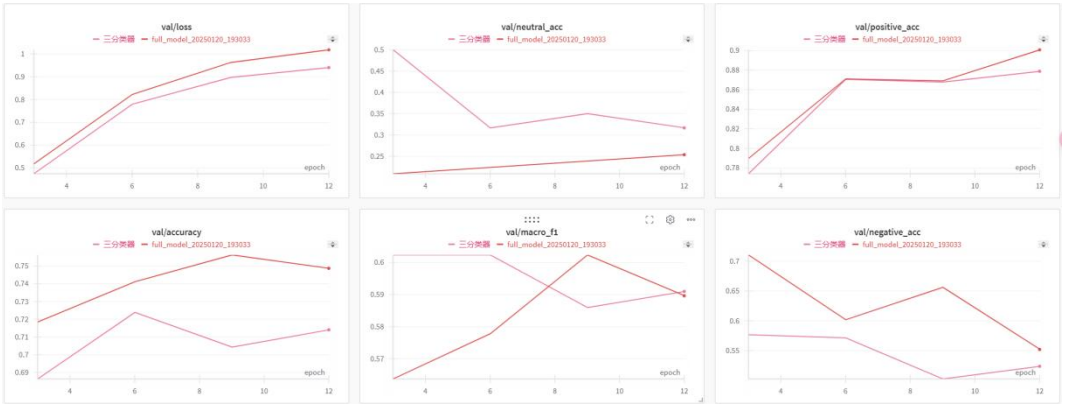
但是事实上效果并不是很好



但是我并没有进行详细地数据增强的超参数搜索，例如增强数据的数量等，也许有合适地超参数可以使得性能得到提升。

2.分类器的设计

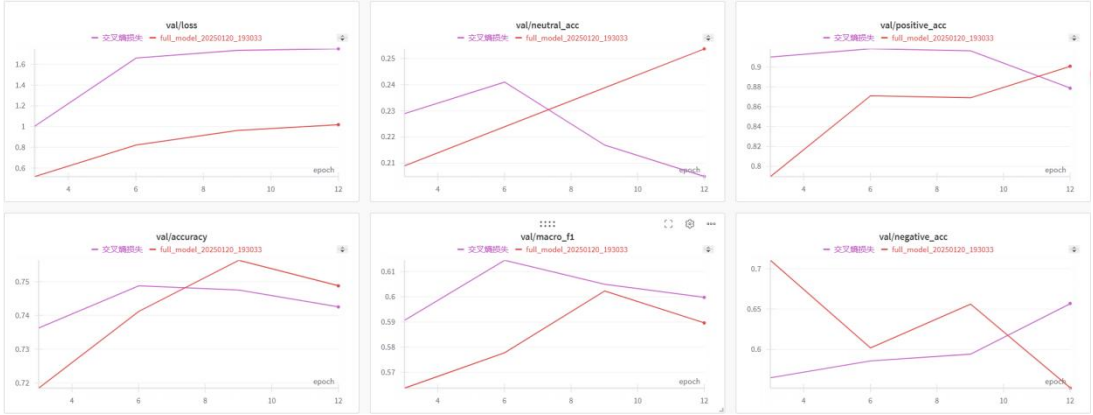
我在进行分类器设计的时候，最初考虑的是设计一个分类器进行三分类，在后面模型设计的尝试中，我拆分了分类器为两个分支，以更好地注意到 neutral 类，但是效果上没有很明显地改善。



3.Focal Loss

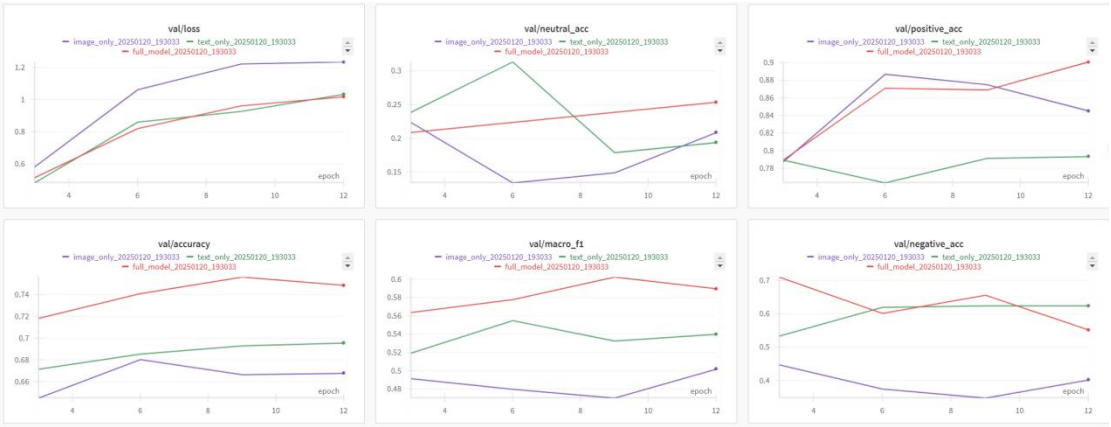
为了应对数据集中不平衡的情况，我还采用了 Focal 损失和交叉熵损失进行组合，以帮助模型提高对于 neutral 样本的区分能力。

以下是混合损失和交叉熵损失的对比，可以看到模型在处理 neutral 类的判断上准确率高了很多



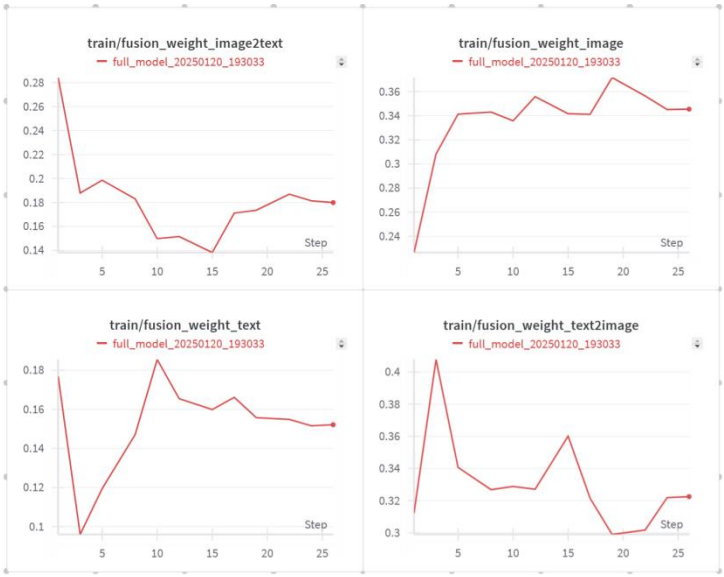
4.多模态的消融实验对比

为了测试我们多模态模型的性能，我还与纯文本、纯图片两种数据训练出的模型进行了对比，数据如下



我们可以发现，我们的多模态模型在准确率与 Macro F1 两种指标上均取得了更好的分数，这验证了多模态融合的有效性。纯文本模型虽然在某些特定情境下表现不错，但由于缺乏直观的图像信息，对于复杂场景的理解能力有限。而纯图片模型虽然能够捕捉到丰富的视觉特征，但往往忽略了文本中的关键信息，导致理解不够深入。相比之下，多模态模型结合了文本和图片的优势，能够更全面、准确地理解输入数据，从而在分类任务中表现出色。

此外，我还对不同模态的权重进行了调整，模型也确实在训练过程中不断调整各类信息的占比以得到更好的结果，下图展示了我们四种特征的权重变化。



六.实验总结

在本次实验中，我成功构建并优化了一个多模态融合模型，用于预测给定文本和图像配对的情感标签。通过分析数据集，我们发现数据存在偏态分布，其中正面情感的样本数量远多于中性和负面情感的样本。为了解决这一不平衡问题，我们采取了多种策略，包括数据增强、使用 Focal Loss 损失函数以及设计专门的分类器结构。

在模型构建方面，我采用了 BERT 模型作为文本编码器，以及 Vision Transformer（ViT）模型作为图像编码器。通过交叉注意力机制，我融合了文本和图像的特征，并动态地调整了不同特征的权重。为了更好地处理中性情感的分类，我设计了一个双分支的分类器结构，其中一个分支专门用于判断中性情感，而另一个分支则用于区分积极和消极情感。

在超参数搜索过程中，我通过验证集对模型进行了细致的调整，并找到了最佳的超参数组合。

在模型训练过程中，我遇到了一些挑战，例如数据增强对模型性能的提升并不总是正面的，这提示我在实际应用中需要仔细选择和调整数据增强的策略。此外，我还尝试了不同的分类器设计，以及结合 Focal Loss 和交叉熵损失的混合损失函数，以提高模型对少数类别的识别能力。

通过与纯文本和纯图像模型的对比实验，我验证了多模态融合模型的有效性。多模态模型在准确率和 Macro F1 指标上均取得了更好的成绩，这表明结合文本和图像信息的模型能够更全面地理解输入数据，从而在情感分类任务中表现更优。

本实验不仅加深了我们对多模态学习的理解，而且通过实践探索了多种技术手段来优化模型性能。尽管在实验过程中遇到了一些挑战，但通过不断的尝试和调整，我成功地构建了一个性能良好的多模态情感分类模型。未来的工作可以进一步探索更有效的数据增强方法、损失函数设计以及模型结构优化，以进一步提升模型的泛化能力和分类准确性。