多模态情感分析实验报告

实验目的：

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，再利用训练好的模型预测测试集（test_without_label.txt）上的情感标签。

方法：

通过预训练模型 BERT 处理文本，ResNet 处理图像，最终通过融合两者特征进行分类。将训练数据划分为训练集与验证集，对模型进行训练，再利用训练好的模型对测试集的情感标签进行预测。

GitHub地址：https://github.com/TATNTA/-

过程：

首先进行数据路径设置与数据加载，通过pandas读取了训练集(train.txt)和测试集(test_without_label.txt)。在这个地方出现了报错“AssertionError:标签值超出范围”，输出中可以看到，train_df['tag_encoded']的唯一值是[3,0,1,2]，这说明标签编码过程中，某个标签被意外地映射到了3，因为只期望得到0，1，2这三类。为此做出如下修改：首先检查标签列中的所有值，确保它们符合预期（即positive,neutral,negative）。如果有其他值（如3），就需要删除或更正；如果发现数据中有额外的无效标签，直接通过isin()方法过滤掉不合法的标签。这样就过滤掉了无效的标签，并保留有效的情感标签：positive、neutral、negative。使用LabelEncoder对标签进行编码，将positive、neutral、negative分别转换为0、1、2，并且保证了标签在[0,1,2]范围内。

接着使用train_test_split将训练集数据划分为训练集（80%）和验证集（20%），然后自定义Dataset类。MultimodalDataset类继承自Dataset，用于加载多模态数据（文本和图像）。通过guid获取相应的.txt文件内容，并使用BertTokenizer进行编码，确保文本输入的长度符合要求（最大长度128）；通过 guid获取相应的.jpg文件，并使用PIL加载图像。如果文件不存在，则使用一个空白图像（224x224）填充。最终，返回的是一个包含文本、图像和标签的元组。

然后进行BERT和图像预处理设置。使用BertTokenizer加载预训练的BERT tokenizer，再使用torchvision.transforms对图像进行处理，调整大小为224x224，将图像转为tensor，并进行标准化。而对于数据加载器，采用的设计是：创建训练集和验证集的DataLoader，设置了批量大小为32，训练集启用shuffle（随机打乱数据顺序）。

接下来进行多模态融合模型设计。使用预训练的BERT模型，从文本数据中提取特征。text_output.pooler_output获取的是文本的全局特征（池化后的输出）；使用预训练的ResNet50模型提取图像特征。这里去除了ResNet的最后一个分类层（self.resnet.fc = nn.Identity()），只保留提取特征的部分。再将文本特征和图像特征按维度拼接，形成一个大的特征向量。最后通过全连接层（self.fc）对拼接后的特征进行分类，输出最终的情感标签。

进行训练和验证，使用交叉熵损失函数（CrossEntropyLoss）进行多类分类任务。使用Adam优化器，设置学习率为。训练过程包括：训练模式，对每个批次计算损失并更新参数；验证模式，计算在验证集上的损失和准确率；早停机制，如果验证集上的损失连续若干轮没有改进，则提前停止训练。

在测试集上进行预测，使用训练好的模型对测试集中的每个样本进行预测，最终将预测结果保存到文件test_predictions.txt。

结果：

选取不同的几组超参数进行了几次训练，结果如下：

超参数		输出结果	
学习率	batch_size	验证损失	准确率
2e-5	16	0.7744	0.7038
1e-5	8	0.7919	0.7125
5e-6	4	0.7344	0.7163
5e-6	32	0.6935	0.7188

由于设有早停机制，因此epoch值取较大值500，等待早停机制发挥作用即可。经过比较可以认为在学习率为5e-6，batch_size为32时训练模型的效果较好，较小的学习率有助于使模型更加稳定地收敛，并且有可能避免之前学习率较大时可能出现的过冲或震荡，此外，选取的批量大小既提高了训练速度，又具有较好的泛化能力。

接下来又进行了消融实验，为了提高效率，选择学习率为5e-6，batch_size为32，epoch = 2，得到结果如下：

训练多模态模型得到结果：Validation Loss: 0.7005, Accuracy: 0.7250

训练文本模型得到的结果：Validation Loss: 1.1339, Accuracy: 0.1163

训练图像模型得到的结果：Validation Loss: 1.0524, Accuracy: 0.4913

可以看出：多模态模型表现出较好的性能，验证损失和准确率明显优于单模态模型。此结果表明，文本和图像的融合提供了有价值的信息，有助于提升模型的识别能力。文本模型的验证准确率较低，接近随机猜测（0.33）。这可能是由于BERT模型在没有图像辅助的情况下无法有效处理特定任务，或者文本数据本身的信息量较少。文本信息可能不具备足够的上下文信息，导致模型无法有效分类。图像模型的验证准确率约为50%，这是一个中等的表现。虽然图像数据本身可能包含了视觉特征，但单独的图像模型仍然无法捕捉到文本中的情感信息，因此准确率仍低于多模态模型。

讨论：

1.**多模态融合的优势：**

优势：多模态模型在验证集上取得了最高的准确率（0.7250）和最低的验证损失（0.7005），表明文本和图像的信息相辅相成，融合效果明显。多模态模型通过结合视觉信息和语言信息，能够提供更多的上下文和情感线索，从而提高了模型的表现。

结果解读：文本和图像的联合学习不仅弥补了单模态模型的不足，还能利用两种模态的信息互补。特别是在情感分析任务中，图像和文本的融合能够从不同的维度理解情感语境，提升模型的综合能力。

2.**文本单模态模型的局限性：**

局限性：文本单模态模型的准确率显著低于多模态模型，且远低于预期。这可能是由于仅依靠文本，缺乏图像这一直观的情感线索，导致模型难以判断情感的准确类别。特别是在情感分析中，图像往往能提供直观的视觉

提示（如表情、场景、颜色等），这些信息可能对于区分情感类别至关重要。

改进方向：提高文本模型的性能可以通过进一步调优BERT模型、增加更多的上下文信息或使用更强的预训练模型。

3.**图像单模态模型的局限性：**

局限性：图像模型的准确率较低（约50%），这表明尽管图像包含了一些有用的信息，但单独使用图像进行情感分析存在一定的局限。图像的情感表达可能不够清晰或与文本不完全一致，尤其是在一些复杂或含糊的情感表达中，单一图像信息可能无法提供足够的辨别能力。

改进方向：在图像单模态模型中，可以考虑引入更多的视觉特征，如使用更深层次的卷积神经网络（例如EfficientNet等）或增加图像预处理步骤，来提升模型在视觉情感识别方面的能力。