《媒体与认知》上机实验最终报告

一、基础部分总结

本次作业的中期，我成功实现了一个简化版的 CLIP 模型，通过图像与文本的对比学习，实现多模态语义对齐，并支持文本与图像之间的相互检索。核心思想是将图像和文本分别编码为向量，并通过 InfoNCE 损失函数在训练阶段拉近正样本对之间的距离，推远负样本对的距离，从而使得模型在共享语义空间中能够理解“图”和“文”的对应关系。

具体而言，中期的基本模型采用双塔结构，实现了使用 ResNet18 进行图像特征提取并投影到嵌入空间的图像编码器，以及使用双层 LSTM 网络提取句子语义信息的文本编码器整个系统的输出为图文嵌入对，使用余弦相似度作为匹配依据，通过双向 InfoNCE 损失函数进行训练，完成了图像与文本嵌入空间对齐的基本目标。在实验过程中，我掌握了对比学习的核心思想、双塔结构的建模方式以及 InfoNCE 损失函数的使用方法。

基础模型在 Flickr8k 数据集上进行了训练和评估，主要观测指标包括损失函数下降曲线以及 Recall@K 检索准确率。

下图展示了训练集、验证集和测试集上的损失下降趋势：

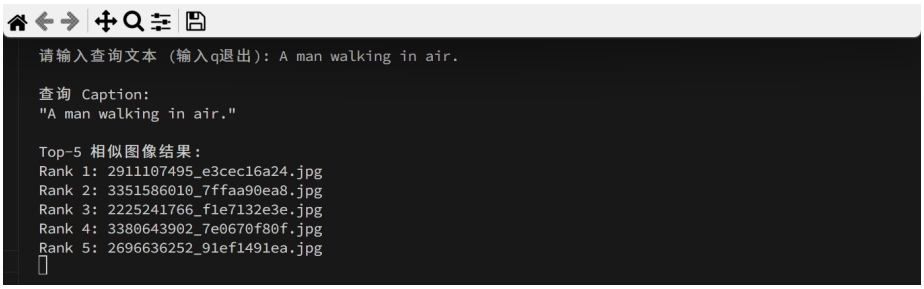
可以看出模型在前几轮收敛较快，之后趋于平稳，表明模型能够一定程度上有效学习图文之间的对应关系。

下图展示了 Recall@K 检索准确率，取 K = (1, 5, 10)：

```
📈 Text → Image Retrieval:
Recall@1: 5.51%
Recall@5: 19.35%
Recall@10: 30.05%

📈 Image → Text Retrieval:
Recall@1: 6.20%
Recall@5: 19.95%
Recall@10: 31.07%
```

除此之外，实现了文本检索图像（Text → Image）的可视化，展示模型对于指定文本返回的 Top-5 图像如下：



由以上数据可见，模型虽然基本学会了图文嵌入对齐，但在高精度检索方面仍存在较大不足。具体表现为：

- Recall@1 偏低，Text→Image 的 Recall@1 仅为 5.51%，Image→Text 的 Recall@1 也只有 6.20%，说明绝大多数检索未命中正确对应项。这可能是由于：
  - ResNet18 和 LSTM 对复杂语义的建模能力有限；
  - 负样本太“容易”，导致模型没学会在相似图文中做精细判别。
- 在文本 “A man walking in air.” 的图像检索中，Top-1、Top-2 命中了滑板空中动作，基本符合句意，但 Top-3 则是一张男人躺在床板上的图像，与句意几乎不符。这可能是由于：
  - 模型更关注显性词汇的共现（如 “man”），而忽略了句子整体语义组合的正确性；
  - LSTM 编码器难以抓住如“in air”这种抽象空间位置关系。

二、关键尝试与改进

1. 换 Transformer

2. 换预训练 BERT

3. 梯度爆炸，加上梯度裁剪、动态学习率、早停

4. 文字和图像数据增强

5. 发现 ResNet18 训练太慢，换预训练 ViT

三、可视化展示

四、总结与反思

本次实验成功实现了一个简化版的 CLIP 模型，完成了图像与文本嵌入空间对齐的基本目标。