代码实现目标

该项目旨在基于 MNIST 手写数字数据集训练一个简单的 CLIP（Contrastive Language-Image Pretraining）模型，用于学习多模态模型。其主要功能包括对图片进行分类以及找出相似图片，且模型仅能处理 0 - 9 的手写数字预测。

主要思路

模型构建：定义CLIP类，包含图像编码器ImgEncoder和文本编码器TextEncoder。图像编码器采用 ResNet 残差网络结构，文本编码器则是简单的嵌入层加全连接层，二者输出的嵌入向量做点积得到 logits。

数据处理：使用torchvision加载 MNIST 数据集，并进行必要的转换。

模型训练：通过交叉熵损失函数计算图像和文本的损失，使用 Adam 优化器进行迭代训练，确保每个批次数据包含 0 - 9 所有数字。

推理演示：加载训练好的模型，对单张图片进行分类，并找出与之最相似的 5 张图片。