题目

摘要

对低分辨率图像的分类任务一直是图像人工智能领域基本的问题之一，CLIP 模型通过从原始文本中学习图像，以及在互联网上收集的庞大图文数据集上预训练，实现了令人瞩目的图像表示学习能力以及图像分类能力。LIIF 模型通过局部隐式图像函数，结合图像坐标和周围 2D 深度特征，实现了对图像的连续表示学习，通过自监督任务，LIIF 模型以高效的方式完成了图像的超分辨率生成，可将连续表示呈现到任意分辨率，甚至高达30倍分辨率。本工作通过在训练 CLIP 模型过程中引入 LIIF 超分模块，将低分辨率图像转换成高分辨率版本，然后使用这些高分辨率图像重新训练 CLIP 模型，从而验证 LIIF 对于 CLIP 执行低分辨率图像分类任务的提升效果。在我们的实验中，引入 LIIF 超分模块的 CLIP 在执行低分辨率图像分类任务时表现出小幅提升。这项研究对于进一步理解超分模块对图像分类任务的影响，以及无监督学习中的潜在应用具有启发意义。

关键词：低分辨率图像；图像分类；CLIP模型；LIIF模型；

1 引言

当进行大规模图像分类任务时，引入 CLIP 模型 [2]连接图像特征空间和文本特征空间展现出了显著的优势。CLIP 模型的独特之处在于其能够有效地实现图像到文本的映射，为跨模态学习提供了一种卓越的解决方案。然而，在应对低分辨率图像分类任务时，CLIP 模型的分类性能可能受到分辨率差异的影响。为了进一步提升 CLIP 模型在低分辨率图像分类任务中的适应性，本研究将超分辨率 LIIF 模块 [1]引入到 CLIP 模型的预训练阶段。通过在图像数据集上进行超分辨率预处理，然后利用这些经过超分处理的图像数据进行 CLIP 模型的预训练，我们旨在增强 CLIP 模型在低分辨率图像分类中的性能。

2 相关工作

2.1 CLIP模型

OpenAI 公司开发的 CLIP 模型标志着图像分类领域的一项重要突破。该模型首次成功实现了图像特征空间和文本语义空间的有机连接。通过使用 4 亿对图像-文本对进行训练，CLIP 通过最大化正样本对之间的相似性，同时最小化负样本对之间的距离，从而为图像和文本之间的对比学习建立了强大的关联。该模型具备了在图像分类任务中实现 Zero-shot 学习的能力，这意味着它没有在下游任务数据集上做额外训练的情况下，仍能够在目标任务的数据集

上表现出卓越的性能。通过输入一张图片，CLIP 能够输出与之最为匹配的文本信息，为图像分类任务提供了一种高效而强大的解决方案。

2.2 超分模型Local Implicit Image Function

目前，2D 图像表示主要采用二维像素阵列，这种方法在复杂性和精度方面受到分辨率的限制，并要求在某一任务中对所有图像进行归一，使其达到统一分辨率。为了克服这些限制，Yinbo Chen 等共同作者提出 LIIF 超分模型，灵感来自于 3D 图像表示方法的隐式神经表达。LIIF 提供了一种连续表示 2D 图像的新方法，通过以图像坐标和相邻的 2D 潜在特征作为输入，从而预测指定坐标处的 RGB 值。值得强调的是，由于 LIIF 考虑到坐标的连续性，它具有出色的图像超分辨率能力，通过这一特性，LIIF 能够在不受分辨率限制的情况下完成任意目标分辨率的图像超分任务。这为解决目前 2D 图像表示面临的分辨率限制问题提供了一种创新性的解决方案，为图像处理任务的高效性和精确性带来了新的可能性。在考虑使用 LIIF 模型进行图像预处理时，其在超分任务中的卓越性能为图像分类任务提供了更为准确和细致的特征，有望在研究领域和实际应用中推动图像处理技术的进一步发展。

3 本文方法

3.1 模型方法概述

CLIP 模型结构如图 1 所示，本模型采用对比学习的方法，使得匹配的“图像-文本”对各自的语义向量拼成特征向量作为正样本，不匹配的图像-文本各自的语义向量拼成特征向量作为负样本，通过优化正负样本之间的损失函数值，正样本之间损失函数最大化（向量空间中尽量靠近），负样本之间损失函数最小化（向量空间中尽量远离），最终实现将正样本的特征样本作为嵌入样本加入到向量空间，此向量空间作为后序测试模型时对比的“先验知识”，达到训练模型的效果，使得测试图像的特征向量能够映射到目标分类的特征向量实现分类。（利用向量空间中的位置远近实现分类）

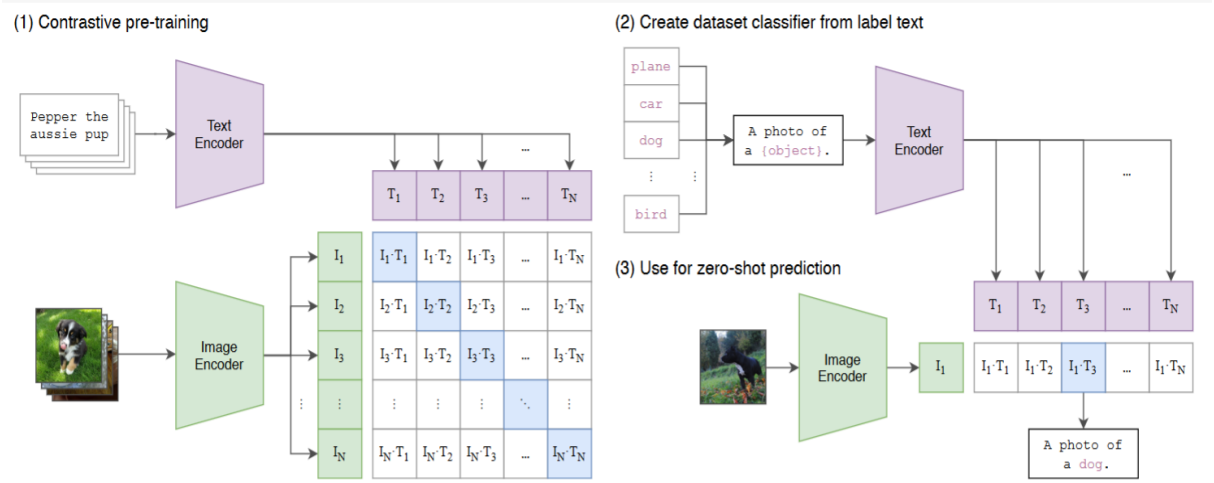

图 1. CLIP 模型图

3.2 对比学习预训练概述

在预训练模型时，模型的输入是若干个“图像-文本”对，如图 1 中 Contrastive pre-training 的输入为一张小狗图片及其文本描述“Pepper the aussie pup”。针对图像输入，模型使用一个 Image Encoder 获取图像特征，这里 Encoder 可以使用 ResNet 或者 Vision Transformer 。针对文本输入，模型使用一个 Text Encoder 获得文本特征。假设每个 training batch 都有 N 个“图像-文本”对，则会有 N 个图像特征 I_i 和 N 个文本特征 T_i 。

获得“图像-文本”特征向量后，使用无监督的对比训练方式，通过交叉熵损失函数计算各图像特征向量与文本特征向量之间的相似性，在通过梯度下降算法最大化主对角线上正样本对之间的相似度。OpenAI 作者团队专门制作一个数据集，其中有 4 亿个“图像-文本”对，且数据清理的比较好，质量比较高。这使得 CLIP 的特征学习效果如此强大。

3.3 Zero-shot任务概述

CLIP 模型因其预训练所使用的训练数据集规模大的优势，采用对比学习链接了图像空间与文本空间，使其能够在下游任务中实现极好的 Zero-shot 能力，即实现仅预训练一个模型，实现在下游任务中不需要再做微调，实现比较好的泛化性。在图 1 中 Use for zero-shot prediction 任务中，利用预训练得到的 Image Encoder 处理目标分类图像即可得到图像特征 I_1 ，利用图像特征向量 I_1 与文本特征向量 T_i 计算 cosine similarity 得到该图像与各文本之间的相似度($T_i \cdot T_i$)，值得一提的是，预训练模型时，原作者并没有设置分类头，这里的文本特征向量 T_i 来自于利用预训练得到的 Text Encoder 去预处理类别句子得到。最终输出 $T_i \cdot T_i$ 最大值对应的类别从而完成该图像分类任务。

3.4 LIIF超分模型

LIIF 模型结构图如图 2 所示，该模型采用两阶段训练结构，第一部分为数据预处理部分，第二部分为训练部分。在第一阶段训练图像通过随机 scale 的下采样生成 Input，通过将训练图像表示为 pixel samples x_{hr}，S_{hr}（其中 x_{hr} 是 pixels 的中心坐标，S_{hr} 是 pixels 对应的 RGB 值）。在第二阶段首先训练一个编码器 E_ϕ。用于将一个图像映射为潜在编码 LIIF 表示，然后使用 x_{hr} 及潜在编码 LIIF，f_θ 预测对应的 RGB 值 S_{pred}，再使用 S_{pred} 与 S_{hr} 计算 L1loss 作为损失，通过梯度下降算法以优化 loss。同时在 batch training 中，从训练集中采样 batches，而 loss 也是所有 instances 的平均值。

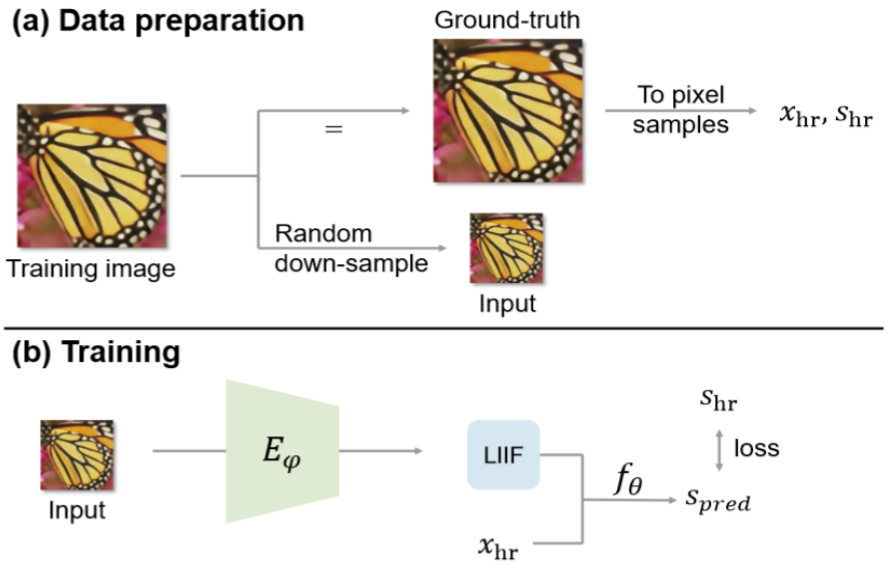

图 2. LIIF 模型图

4 复现细节

4.1 与已有开源代码对比

本文工作参考 OpenAI 公司提供的开源代码，在训练阶段引入超分模型 LIIF 对图像进行超分预处理，如图 3 所示。本实验旨在研究在低分辨率图像分类任务中，使用在超分数据集上预训练过的 CLIP 模型是否对分类性能有提升，该实验对于分辨早期设备产生的低分辨率图像具有指导意义，有助于指导相关工作者选择合适的图像大模型进行图像分类任务。

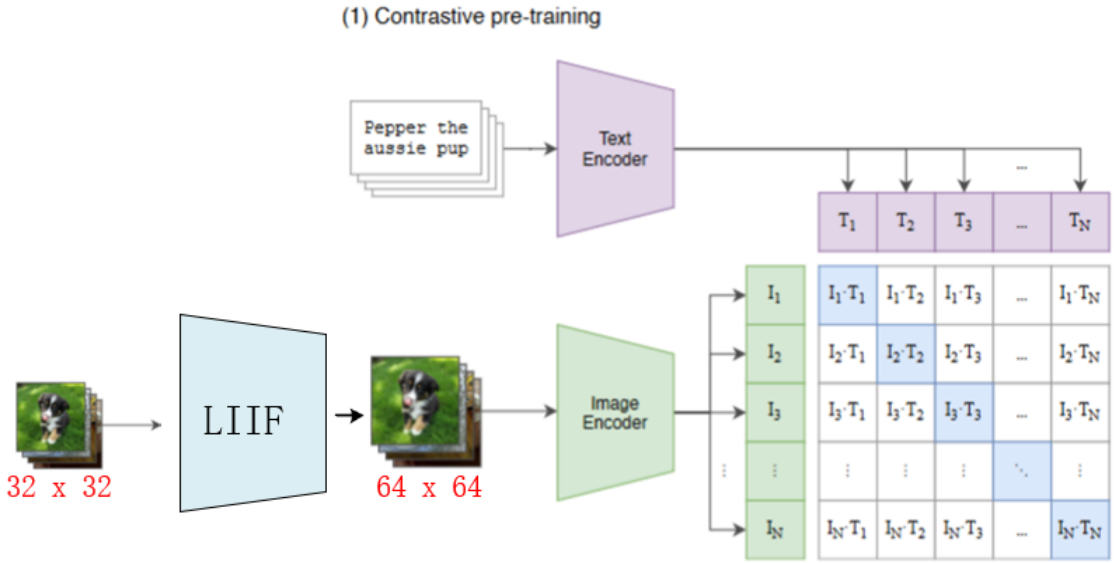

图 3. 引入LIIF后的CLIP模型图

4.2 数据集准备

本实验取自 cifar10 数据集，该数据集由两部分组成，一部分为图像集，一部分为“图像-类别号”的映射文件，本实验采用离线下载的方式，将所有训练集、测试集图片分别解压缩得到两个存放 jpg 图片的文件夹，然后使用 Python 代码制作本实验使用的 my_train_cifar10 训练集和 my_test_cifar10 测试集。

获得的训练数据集 my_train_cifar10 来自 cifar10 中训练集的 1 万张 32 x 32 低分辨率图片，测试数据集 my_test_cifar10 来自cifar10中测试集的 2 千张 32 x 32 低分辨率图片。实验方法采用对照实验，对照组为使用 CLIP 模型在原分辨率数据集 my_train_cifar10 上直接训练，实验组为使用引入 LIIF 超分模块的 CLIP 模型在超分辨率后数据集 my_train_cifar10_afterLIIF_x2 上训练，两次训练的训练迭代相关参数如表 1 所示。

表 1. 训练实验参数

实验组别	分辨率	训练集图片规模	epoches	batch_size	lr
对照组	**32 x 32**	1万张	25	32	1e-6
实验组	**64 x 64**	1万张	25	32	1e-6

4.3 实验环境搭建

本实验使用 Google 公司提供的可在线编写和执行任意 Python 代码的平台 Colab，以及云存储平台 Google drive 以克服计算资源有限的问题。

4.4 界面分析与使用说明

实验平台如图 4 所示，需要注意的是 Colab 为不同用户提供了不同的计算资源，本实验采用其提供的 V100 GPU 训练 CLIP 模型。同时需要注意的是 Colab 自动配置了最新的 Cuda 环境，而每次的连接云端服务器都会重制上次连接手动下载的一切包，为此特别设置下载环境所需要的包的模块。

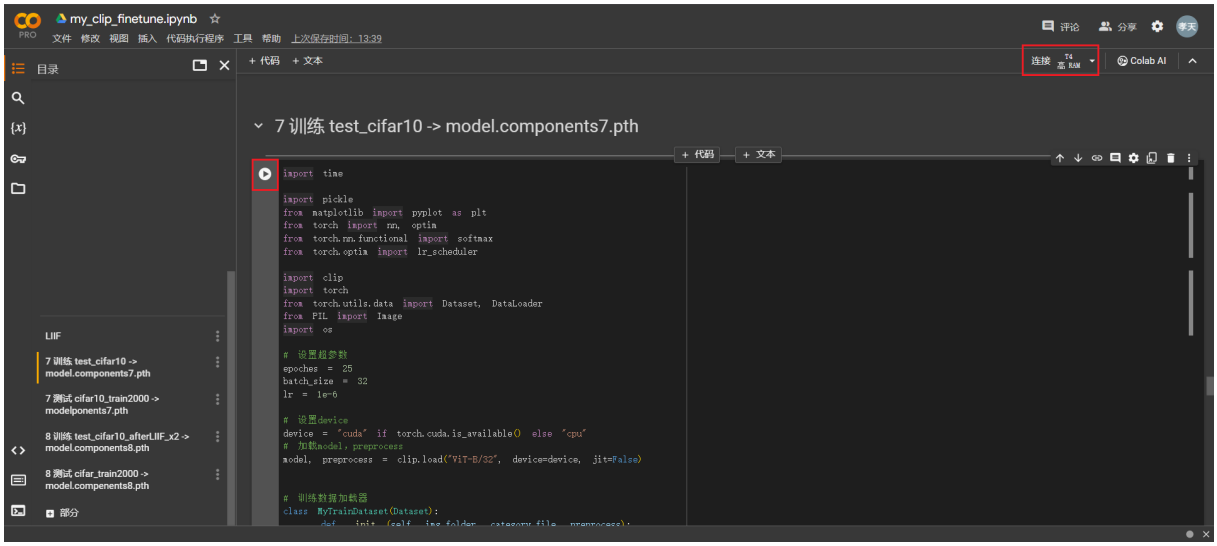

图 4. 操作界面示意

4.5 创新点

本实验在低分辨率图像分类任务中，预训练引入超分模块的 CLIP 大模型以实现高质量的分类结果。引入超分模块使得 CLIP 可以学习到图像的超分先验，使得预训练得到的模型执行低分辨率图像分类任务时能够提升其分类正确率。

5 实验结果分析

本实验研究引入超分模块对 CLIP 模型在低分辨率图像分类任务上的正确率提升效果，因此采用对照实验见表 2 所示两组实验参数。在测试环节，将两次预训练得到的模型分别在同一个测试集 my_test_cifar10 上测试其正确率，测试结果如表 2 所示，可以看到实验组的正确率有 1.5% 的提升，由此可见，引入超分模块对于 CLIP 在低分辨率数据集上的图像分类效果有所提升。

表 2. 实验结果

实验组别	训练集分辨率	测试集	测试集分辨率	正确率
对照组	**32 x 32**	test_cifar10	32 x 32	**95%**
实验组	**64 x 64**	test_cifar10	32 x 32	**96.5%**

6 总结与展望

本实验通过对照实验的方法，分别测试了原生 CLIP 模型和引入超分模块的 CLIP 模型分别在低分辨率图像数据集上的分类效果，实验结果说明，在预训练阶段引入超分模块 LIIF 得到的预训练模型 CLIP 对于低分辨率图像分类效果有所提升。但是，实验存在很大的局限性，主要体现在数据集规模比较小，种类比较少，需要进一步的在大数据集上测试其效果。同时目前引入超分模型的方法只是图像预处理部分，需要进一步思考如何将 LIIF 嵌入到 CLIP 模型之中以实现一键式的在低分辨率数据集上实现先超分在预训练模型。本实验结果的提升效果说明，本实现提出的方法利用 CLIP 模型在超分数据集上预训练有助于提升 CLIP 模型在低分辨率图像分类任务中的正确率。

参考文献

[1] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.