基于MAE的预训练方法

太难实现。。。

参考文献

- MAE : **Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, Ross Girshick. "Masked Autoencoders Are Scalable Vision Learners." arXiv:2111.06377 [cs.CV], Submitted on 11 Nov 2021 (v1), last revised 19 Dec 2021 (this version, v3).**

数据准备

准备无标签的图片数据 从ImageNet等较大数据集获取

超参设置

- Masked Patch Rate： 75% （较高的遮盖率能够促使模型去学习读取特征）
- 基于ViT的框架
  - 参数设置见 **Image-Pre-train Survey.pdf**

模型结构

- 非对称的Encoder 和Decoder
  - encoder
    - 基于 Vit
    - Only Unmasked patch will be encoding
    - **masked patch is ashared, learned vector that indicates the presence of a missing patch to be predicted. We add positional embeddings to all tokens**
  - decoder
    - 较小的架构 计算开销小
    - 最后一层是线性层 投影到patch像素数量的长度 reshape后得到结果

训练目标函数

decoder输出的内容与原先被遮蔽的内容进行比较，直接做像素上的比较
MSE作为损失函数
反传梯度 达到优化的效果

评估

-
  - 微调适应下游任务，例如在不同数据集上的分类任务

    - 把上述任务的结果与传统CNN、ViT在相同任务的结果进行比较
- 冻住主体的ViT（不改变特征提取能力），只学习最后的全连接层，在下游任务上测试效果