Masked Autoencoders Are Scalable Vision Learners

呜呜呜隔壁又有兄弟脱单了

Abstract

之前的ViT作者提了一嘴“要探索视觉中Transformer无监督学习”，这不咱们的MAE就来了嘛。

MAE主要的特点是

1）使用了非对称结构，简单来说就是编码器和解码器输入成分是不一样的（解码器多了mask的图像部份）

2）输入图像的mask比例大大提高到75%，进行了数据增强

MAE的主要特点不是准确率高，而是训练非常快（对于meta来说啦，我们这种小作坊还是玩不起的）

Introduction

在NLP任务中，BERT结构取得了非常好的效果。于是视觉任务也想学习BERT使用掩码自编码器。但是在真正实验时，研究者发现了图像和文字的不同之处，这导致无监督学习的应用寸步难行。

1）网络架构问题。在过去的视觉任务中，CNN占主导地位，掩码和位置编码很难被添加进去，不过这问题已经被ViT解决了

2）信息密度问题。不同于语言文字的高密度语义信息（BERT做完形填空时如果多mask几个词，就不知道这句子想干什么了），图像的信息冗余往往非常大（各位小黑子化成灰都能这认出我家鸽鸽，虽然我觉得这更多是渲染问题）。

3）图像和语言映射后的潜在语义问题。图像的解码器是将像素重组（我也没太懂），因此语义级别比较低；而语言的潜在语义级别很高。因此解码器器在图像处理中非常重要，而处理语言的MLP则显得微不足道。（maybe？）

为了解决上述问题，首先本文大大提高了掩码比例，随机掩盖了图像大约70%的部份，是某种意义上的数据增强，给模型上压力，迫使模型学习更多图片的特征。

除此以外，MAE还使用了非对称结构，encoder仅仅输入可视的图像patch，而在decoder中，mask部份则被处理为同一向量输入。这样的非对称架构大大减少了计算量。

Approach

Masking：首先像ViT所做的那样，将图像分成小patch，再利用随机抽样抽出一部分patch（大概25%），剩下的全部mask掉。这样模型就不能轻易地通过对相邻像素进行插值而得到原图像，迫使模型进行特征学习。

MAE encoder：结构基本和ViT相似，但是不输入mask部份以减少计算量。

MAE decoder：不同于ViT，MAE的解码器只作用于预训练进行图像重建。所以就不用管编码器的大小和结构了，因此解码器可以非常轻量化。今天编码器不在家哦～同时，被mask的图像patch会被共享为同一个可学习向量，一起输入decoder。

Others：MAE采用MSE损失函数，也会和ViT一样输入位置编码。

Over，以后有空了再回来写实验部份。