**论文阅读报告**



**2022 至2023 学年第 二 学期**

课 程 名 深度学习与大数据智能

学生学号 20204205

学生姓名 黄昊

任课教师 文静

报告得分

**《Masked Autoencoders Are Scalable Vision Learners》阅读报告**

**摘要：**本篇阅读报告介绍了Kaiming He等人于2022年在CVPR上发表的工作——Masked Autoencoder.该工作所提出的算法十分简单：对图像进行遮掩操作，然后构建了基于ViT的非对称的编码器-解码器架构对有遮挡的输入图像进行重建任务。本篇阅读报告详细介绍了该工作所使用的方法，分析了可能的动机和优点，以及该方法在一些大型数据集上的表现。

**关键词：**Masked Autoencoder, ViT, Computer Vision

**一、阅读报告简介**

这一小节介绍了作者所选择论文的基本内容，选读原因和阅读本论文时所使用的方法。

**1.1.论文简介**

这篇论文是[1]由Kaiming He等人于2022年在CVPR上发表的一篇工作。这篇论文跟进了ViT[2]的工作，基于ViT提出了Masked Autoencoder架构。这篇文章的特点是对图像的多个patch进行遮挡，并构建了非对称的编码器-解码器架构来解决被遮挡的图像的重建操作。

**1.2.选读原因**

ViT论证了在NLP中常用的Transformer在CV任务中同样可以工作得很好。而MAE跟进了ViT的工作，同时作者也是大名鼎鼎的残差神经网络ResNet的发明人。这两点理由是作者选读该论文的理由。

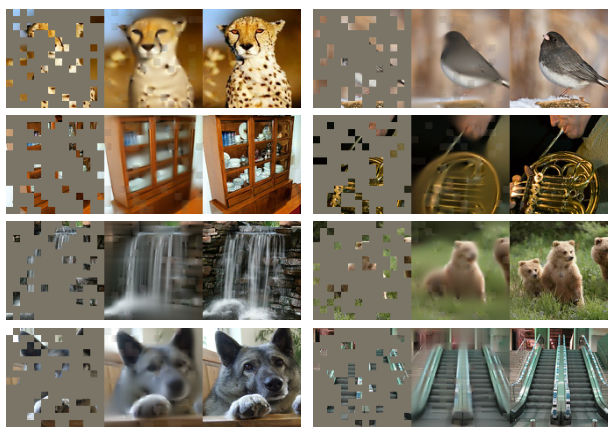
**1.3.阅读方法**

作者在阅读本论文时，首先关注本篇论文的摘要，引言和结论，从整体上把握本篇文章所提出的方法，主要思想，应用与缺点；然后仔细阅读本文的方法部分，从细节上把握本篇文章所提出的模型架构；最后阅读本篇文章所做的实验，从一系列实验中进一步了解本篇文章所提出架构的优势。

**二、论文介绍**

**2.1 论文背景**

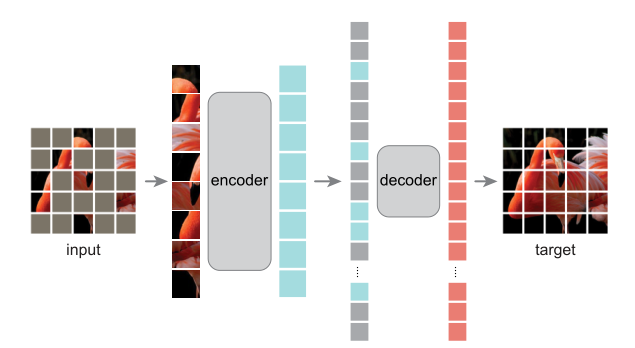
ViT的出现吸引了一些学者研究Transformer架构在CV领域的应用。而Kaiming He等人在去年发表了Masked Autoencoder(MAE)，该架构主要利用基于ViT的非对称编码器-解码器架构，解决带遮挡的图像重建的问题（如图1所示）。



**图1 图像重建（最左边为被遮挡的原图，最右边是原图，中间是还原生成的图片）**

在遮挡比例高达80%的情况下，该架构还能还原出原图的大体样貌。虽然在细节上和清晰度上有一定不足，但在这么高的比例下，还能还原出原图的大体样貌，足以说明该工作所提出的架构还是有一定的价值的。

**2.2 论文所用方法及分析**



**图2 Masked Autoencoders架构的基本流程**

论文在第三小节详细介绍了Masked Autoencoders（MAE）是如何工作的，整体架构如图2所示。具体步骤如下所示：

第一步，首先对数据进行处理。本文所采取的处理方式是对图像进行遮挡(Mask)操作。具体地说，这一步的作用是将原始输入划分为多个patch，然后依从均匀分布随机挑选出一部分patch作为MAE的输入。总的来说，遮挡操作使得被遮挡的patch不再成为模型的输入，从而使得输入变得高度稀疏，大幅度降低了后续所构建模型的复杂度。其次，服从均匀分布的随机采样防止了可能存在的中心偏差（即图像中心存在更多被遮挡的patch）。

第二步，开始构建MAE的编码器。作者介绍编码器是一个ViT[2]。这里先介绍一下ViT：

**（介绍ViT）**

介绍完ViT后，MAE的编码器的结构也就清楚了。最后需要注意的是，Decoder的输入不包括被遮挡的Patch。这一操作大大降低了模型复杂度，这意味着需要使用的内存大大减少，使得模型规模能够有效地减少。

第三步是构建MAE的解码器。解码器的输入是一系列的token，这一系列token包括patch送入编码器后产生的输出，以及一系列的被遮挡的patch对应的token。每个token是一个像素向量，用来表示需要被预测的patch（也就是在数据预处理过程中被遮挡的patch）。这些token还需要加上一系列的位置嵌入，来表示这个token在原来的图片中代表哪个位置的patch。而解码器又是另外一系列的Transformer，用来在预训练期间执行图像重建任务。另外，解码器可以独立于编码器而灵活地进行设计，使得解码器可以使用比编码器规模更小的架构，也就是允许解码器采用非对称的设计，使得训练时间大幅度缩短。

最后就可以使用整个模型对整张图片进行重建了。解码器的输出的每一个元素相当于原有图片对应位置的patch的像素向量。输出通道数量与patch数量相等。损失函数为所有被遮挡的patch复原后与原有patch的像素的均方误差。这里不计算未被遮挡的patch的原因是，其输出和输入都是相同的，对未被遮挡的patch计算均方误差损失没有意义。此外，作者表明如果对像素进行归一化的预处理，就能提高图片的生成质量。

总体来说，这篇文章提出的算法比较简单。简单归纳一下，这篇文章所要解决的问题是对一张具有遮挡的图像进行恢复。其提出了非对称的编码器-解码器架构，使得解码器架构的设计有轻量化的可能；同时送入编码器的图像不包括被遮挡的patch。这两处关键设计使得该模型的训练速度得到一定程度的提升，模型大小能够得到缩小。实际上，MAE对图片分多个patch，设置mask的设计类似于NLP中BERT模型[3]对输入token的mask操作，因此MAE所完成的任务是一个自监督任务。回想一下之前在编码器部分中所提到的ViT模型，其可以认为是视觉任务上的Transformer（Vision transformer），而BERT采用了Transformer的同时，也具有mask操作。可以说，MAE对多个patch的遮挡操作与BERT中对输入token的mask操作的思想很类似，可以认为作者从这里得到了灵感启发。

**2.3 论文实验结果**

本文作者进行了大量实验，总结和分析如下：

**三、总结与个人感悟**

**总体来说，**

**参考文献：**

1. He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 16000-16009.
2. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.
3. Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

正文小标题采用四号宋体，正文部分采用宋体小四号，段前0.5行，行间距20磅，主要内容需包括选读论文简介，选读原因，阅读方法，论文背景，论文所用方法（需分析采用该方法的原因），论文实验结果，总结与个人感悟