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(Mask)，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。

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。

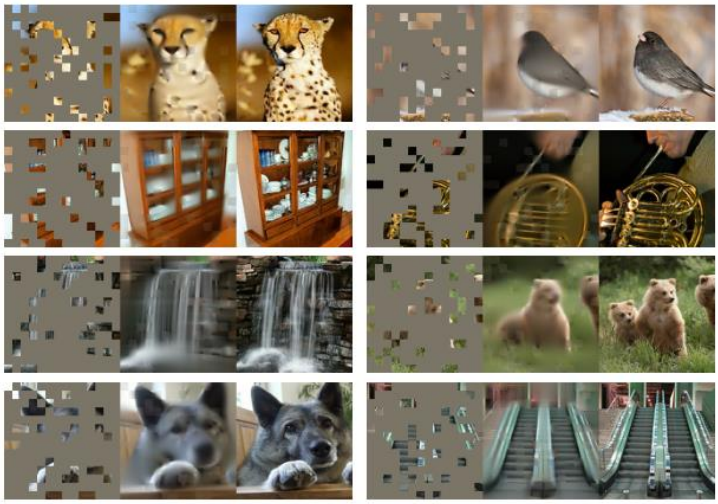

图 1 图像重建（最左边为被遮挡的原图，最右边是原图，中间是还原生成的图像）

2.2 论文所用方法及分析

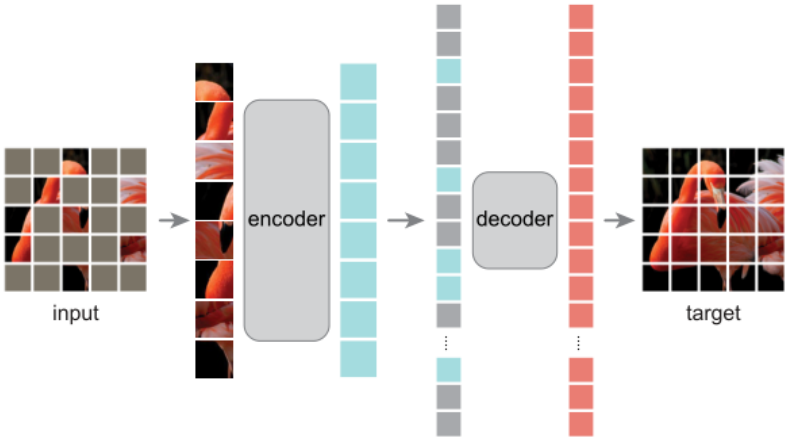

图 2 Masked Autoencoders 架构的基本流程

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[1]。这里先介绍一下 ViT，如图 3 所示。

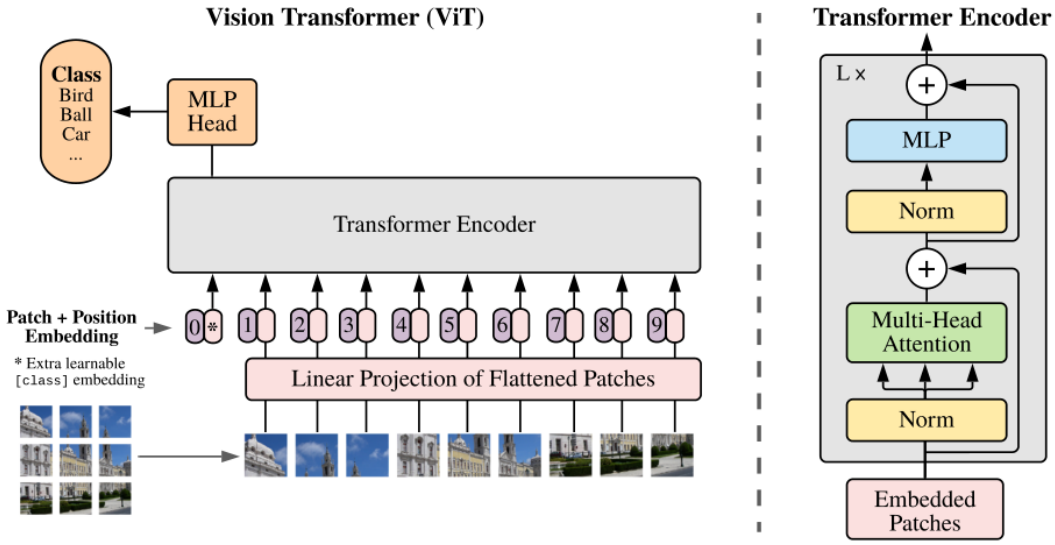

图 3 ViT 的基本架构

ViT 的基本思想是利用 NLP 任务中的 Transformer 的做法，将图像分成多个 Patch 并排成一列，线性投射后得到一个特征，做位置嵌入后送入 Transformer 编码器中。需要注意的是，为了完成分类任务，进入 Transformer 的输入还有可学习分类嵌入(learnable class embedding)的输入。最后有一个全连接层，用以得到最终的分类。

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 在原来的图像中代表哪个位置的 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像进行重建了。解码器的输出的每一个元素相当于原有图像对应位置的 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，就能提高图像的生成质量。

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 对图像分多个 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

本文作者在 ImageNet-1K 数据集上进行了一系列实验，总结和分析如下。

2.3.1 遮挡图像比例对结果的影响

图 4 展示了遮挡比与验证集准确率的关系。可以发现，遮挡比高达 75%的情况下，对于线性探查和微调的情况下效果都不错。并且这个遮挡比例比类似的工作所展示的遮挡比更高。

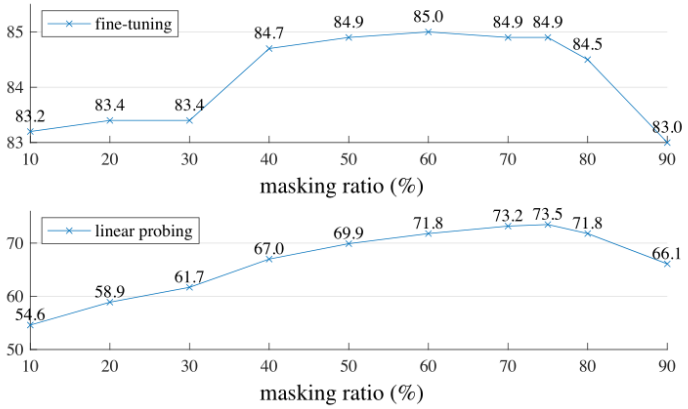

图 4 遮挡比例与验证集准确率的关系

2.3.2 解码器设计对结果的影响

由于在前面提过，解码器的设计是灵活的，因此作者对于解码器的transformer块的数量和通道数不同情况下的性能进行了试验。表1和表2分别反映了不同transformer块和不同通道数下的实验结果。

表1 不同transformer块的结果

blocks	ft	lin
1	84.8	65.5
2	**84.9**	70.0
4	**84.9**	71.9
8	**84.9**	**73.5**
12	84.4	73.3

表2 不同通道数下的结果

dim	ft	lin
128	**84.9**	69.1
256	84.8	71.3
512	**84.9**	**73.5**
768	84.4	73.1
1024	84.3	73.1

其中表1改变的是transformer块的数量，表2改变的是通道数量。从表1可以看到，对于线性探测的方法来说，适当地增加block的数量可以提高一定幅度的效果，但对微调的影响较小；从表2可以看到，通道数的改变结果的改变不如块数的改变明显，其中在512个通道下的效果相对来说有更好的表现。

2.3.3 编码器是否使用被遮挡的token对结果的影响

本文在设计MAE的编码器的时候，未使用带遮挡的patch进行输入。本文对是否使用带遮挡的patch也进行了实验，实验结果如表3所示。

表3 是否使用带遮挡的patch下的结果

case	ft	lin	FLOPs
encoder w/ [M]	84.2	59.6	3.3×
encoder w/o [M]	**84.9**	**73.5**	**1**×

从结果可以看出，对于线性探测来说，使用带遮挡的patch的效果更差，精度约下降14%。作者解释这可能是由于这些被遮挡的patch在未损坏的图像中的相应位置不存在，这可能导致模型部署的时候，准确率会降低。因此，该模型不把被遮挡的patch作为编码器的输入，一是可以加快训练速度，减小训练成本，二是可以提高模型的准确性。

2.3.4 重建图像的处理对结果的影响

表 4 对重建图像的处理的精确度结果

case	ft	lin
pixel (w/o norm)	84.9	73.5
pixel (w/ norm)	**85.4**	**73.9**
PCA	84.6	72.3
dVAE token	85.3	71.6

如表 4 所示，作者对重建目标采取了一系列的处理手段：对 patch 进行归一化；对像素进行归一化和对 patch 进行 PCA 降维。从结果来看，使用归一化后的像素有相对更好的结果。另外，作者还采用了 DALLE 预训练的 dVAE 作为标记，结果显示效果提升不大。因此，对像素的归一化在实验中是最好的方式。

2.3.5 数据增广方法对结果的影响

在处理图像的任务中，我们常常需要对输入进行预处理。作者对一系列的预处理手段进行了实验，如表 5 所示。

表 5 使用不同的数据增广方法的精确度结果

case	ft	lin
none	84.0	65.7
crop, fixed size	84.7	73.1
crop, rand size	**84.9**	**73.5**
crop + color jit	84.3	71.9

从表中看到，随机裁剪的效果较好，无论是固定大小还是随机大小。但使用颜色抖动的预处理效果较差，但值得注意的是，不用任何数据增强的方法，结果也不差，其原因是每一轮模型的迭代，都会选取不同的 patch 进行遮挡。从这一点意义上说，就已经做了很多数据增强的操作了。所以不用其他的数据增强的方式，效果也不差。

2.3.6 遮挡方法对结果的影响

作者在原始模型中采用了随机采样的方法，给不同的 patch 进行遮挡。作者在这个实验中尝试了其他不同的遮挡方法：随机遮挡；大块遮挡以及格点遮挡。如图 5 所示。

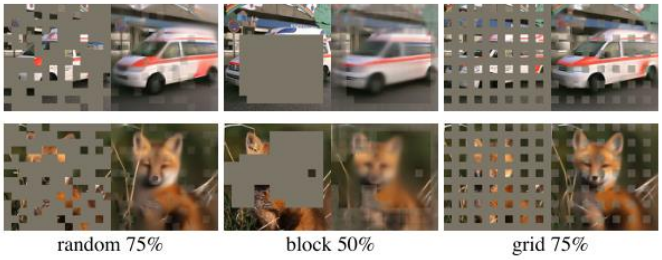

图 5 不同遮挡策略的示意图

可以看到，随机遮挡复原的效果相对来说比较清晰，但是有一定的斑块，大块遮挡的还原效果比较模糊，但值得注意的是，未被遮挡的部分有棋盘的样子，但被遮挡的部分没有，只是比较模糊；而格点遮挡的图像，在还原结果中，有相当规律的灰色块。复原图像出现棋盘式的 patch 的原因可能与训练方式有影响：每次都会采取不同的 patch 被遮挡，其结果是在未被遮挡的 patch 也可能进行了某种变换，使得有这种棋盘状的斑块出现。而准确率结果如表 6 所示。

表 6 使用不同遮挡策略的准确率结果

case	ratio	ft	lin
random	75	**84.9**	**73.5**
block	50	83.9	72.3
block	75	82.8	63.9
grid	75	84.0	66.0

从结果上看，随机采样的效果相对来说也是最好的。同时，随机采样也提供了较高的遮挡比例。总的来说，随机采样是一个较好的策略。

三、总结与个人感悟

3.1 模型总结

总的来说，这篇论文所提出的模型算是 ViT 模型提出后所跟进的工作。其核心思想有二：一是借鉴 Bert 对数据的处理，对图像采用遮挡处理，并只使用不带遮挡的 patch 作为模型的输入，大大减少了训练成本；二是采用非对称的编码器解码器结构，使得模型具有良好的可扩展性质，对于解码器来说，其可以采取轻量化设计，可以进一步地缩减模型规模。需要注意的是，对于图像的处理，需要依从随机分布来选取需要遮挡的 patch；由于每轮训练都会选取不同的 patch，因此数据增广仅采用简单的裁剪操作就可以获得良好的效果。

3.2 个人感悟

这次阅读算是体验了一下跟进前沿研究的情况。实际上，跟进前沿研究的难度没有想象的那么大，主要是因为阅读这些论文不见得需要你新学习多少东西，阅读这些论文需要的知识实际上在课程上都学到了，而不会因为要阅读这些论文多学习啥东西，付出巨量的学习成本。而且，这些论文提出的思想也不一定很高深，阅读下来也不一定有多少难度。甚至这篇论文所提出的思想很简单，如果对 Transformer 和 BERT 熟悉的话，结合 ViT 的工作，就会感觉这篇文章所提出的想法很符合直觉，但从结果来说也很好。这让我也体会到了，很多研究提出的方法也不一定复杂，但是结果很好，这也许是做科研的一种魅力吧，简单，但有效。

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.