**Title:** Masked Autoencoders Are Scalable Vision Learners

**Abstract:**

MAE是可扩展的自监督学习器，本文的MAE方法: 随机掩盖输入图像的一些块，随后重建丢失的像素。

MAE包括两个核心设计：一是只运行在没有mask标记的图像块上进行编码，二是一个轻量级的解码器用于重建缺失的像素。

作用：加速训练和提升精度，例如，一个普通的ViT-Huge模型仅在ImageNet-1K数据上达到了它的最佳精度（87.8%）

**Intro：**

深度学习得益于模型的复杂和硬件的发展，现在往往需要上亿张标签图片，此问题已经在NLP领域中被解决—通过自监督的预训练，比如自回归的语言模型GPT和在BERT中的mask autoencoder：移除了一部分数据并学会预测移除的数据。

BERT成功之后，在掩盖自动编码器这块CV和NLP还有很大差距，是什么让MAE在视觉和语言中不同？本文从以下三个方面剖析：

（i）结构上：CNN占据主导的CV领域不能直接运用这项技术，但随这transformer的引入，不再成为障碍

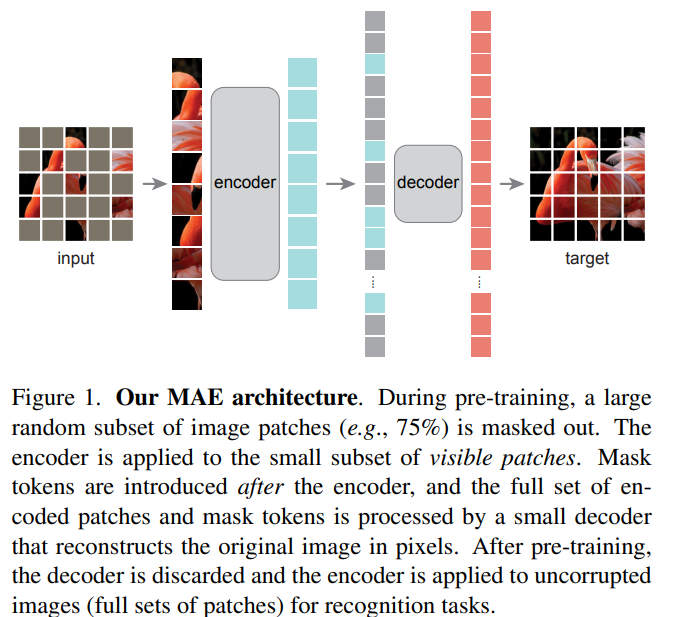
（ii）信息密度：语言是人造的信号，缺失很少的部分会造成很大困难的语义理解；但是图像是自然的信号，具有高度空间冗余信息。解决办法：高比例的对图片进行掩盖，使其成为有挑战的自监督学习任务。

（iii）MEA的解码器：解码器在视觉和语言中做的事情不同，在语言中缺失的token具有丰富的语义信息，而图像则不然，在BERT中它是可被训练的MLP，本文发现解码器的设计决定了隐层表示的语义等级

基于上述分析，本文提出了MAE for CV，设计了非对称的编码-解码器，通过随机掩盖输入的块（eg. 75%），致使训练时间大幅减少（3X或者更多），使之能推广到更大的模型上。

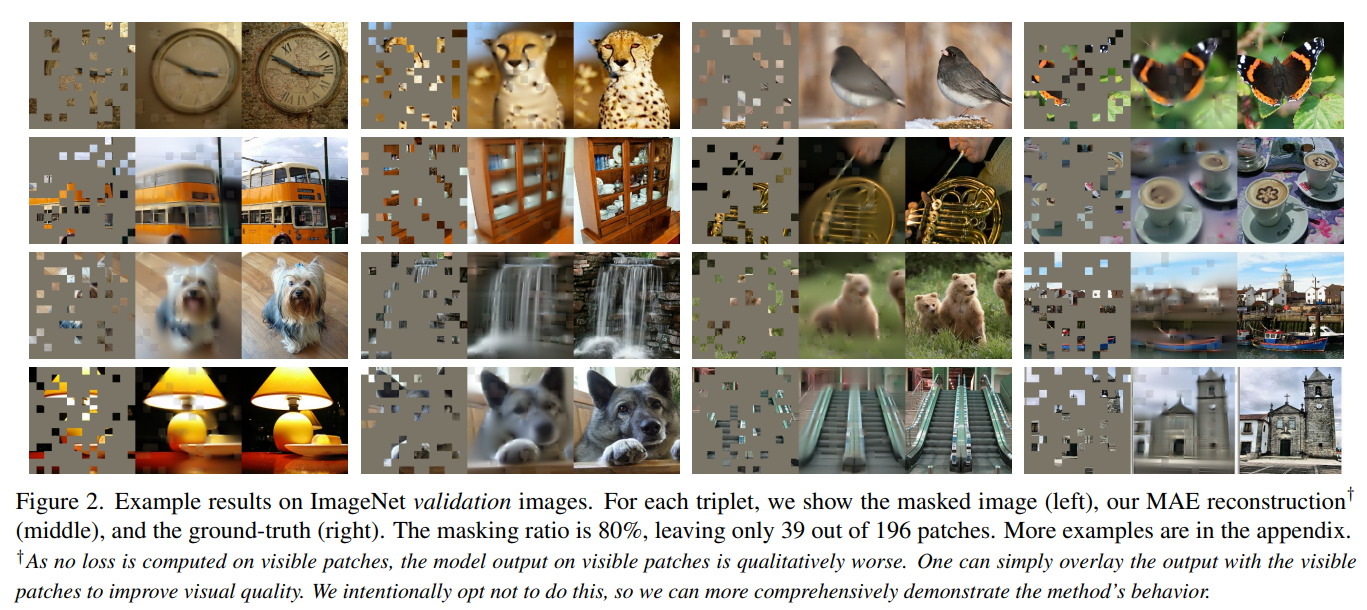
MAE可以学习更大容量更高泛化性的模型，训练ViT-Large/Huge仅在ImageNet-1K上微调，就取得了最好成绩（87.8%），本文还评估了物体检测、实例分割和语义分割等任务的迁移学习。在这些任务中，预训练取得了比监督预训练更好的结果，本文还观察到通过扩大模型规模可以显著提高性能。这些观察结果与自然语言处理（NLP）中的自监督预训练观察到的结果一致

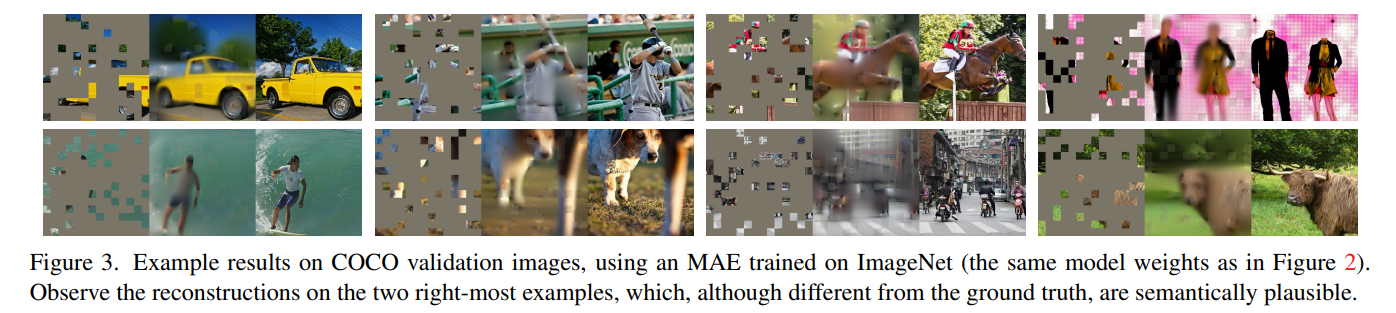
**Figure：**

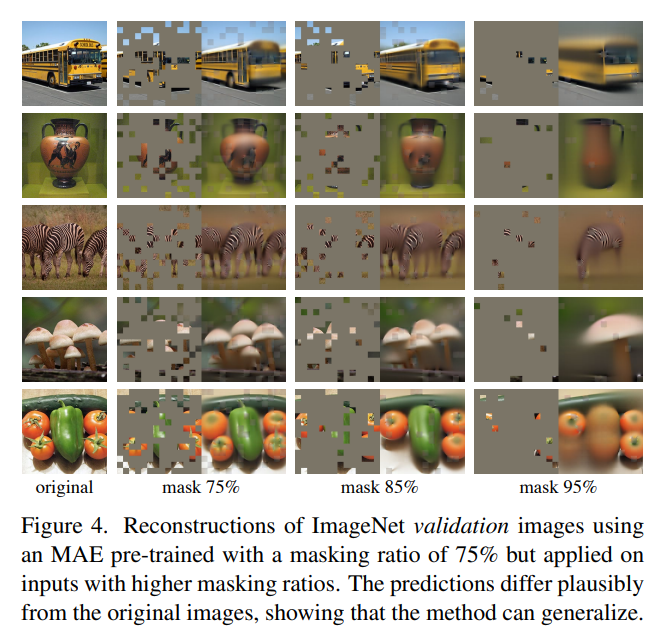
MAE的结构

75%的输入图像块被掩盖，编码器只作用在25%的块上，掩盖的块和被编码器作用的块一起被送入更小的解码器内，编码器重建原始像素。在预训练之后，解码器被丢弃，而编码器被应用于图像识别任务

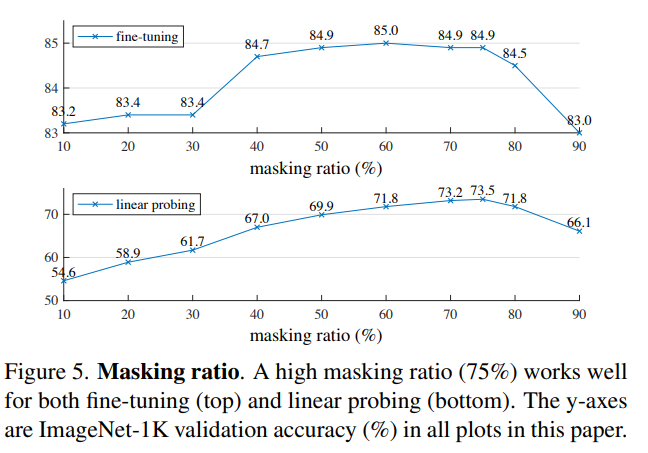
相当于做pre-train

在ImageNet验证图片上的示例结果（80%masked）

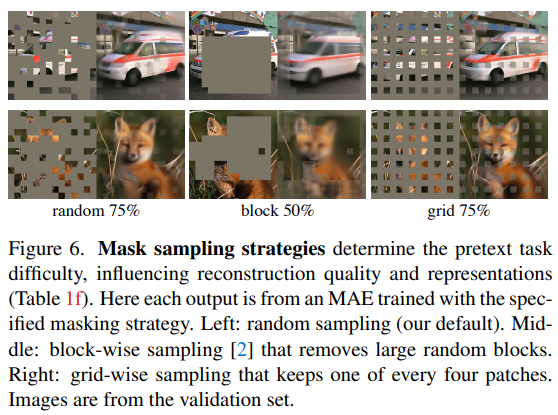
在COCO验证图片上的示例结果，使用MAE在ImageNet上的预训练模型（和Figure2一样），观察最右边两个图片，虽然与事实不同，但语义上是合理的



使用75%掩码率进行预训练的MAE对ImageNet验证图像进行重构，但是对输入图像应用了更高的掩码率。预测结果与原始图像有差异，这表明该方法可以泛化。



掩码率。对于微调和线性探查（从上到下），高掩码率（75%）的效果很好。在本文的所有图表中，y轴均为ImageNet-1K验证集准确率（%）。

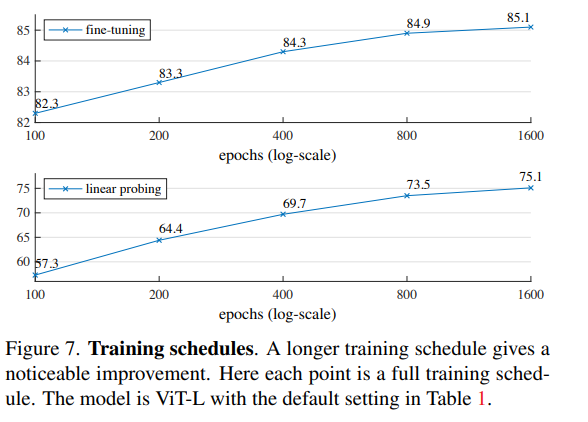


掩码采样策略决定了预处理任务的难度，它会影响重建质量和表示。这里的每个输出都是使用指定的掩码策略训练的MAE模型得到的。

左图：随机采样（我们的默认采样方式）。

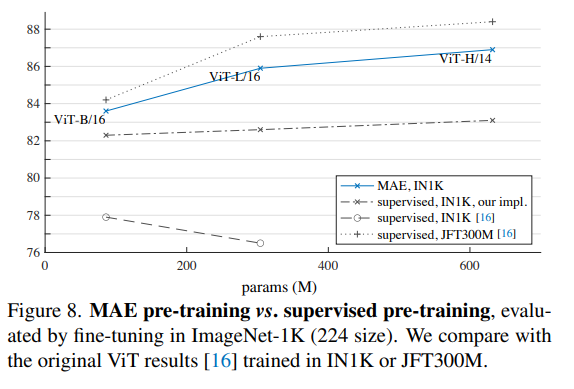
中图：块采样，该策略随机移除大块像素。

右图：网格采样，该策略每四个补丁保留一个。图片均来自验证集。



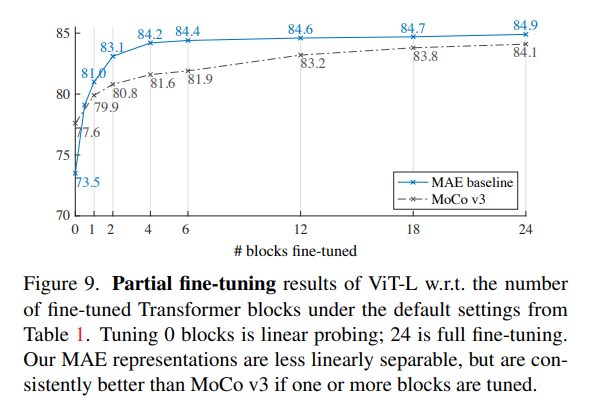
训练计划。更长的训练计划可带来显著的提升。这里的每个点代表一个完整的训练计划。

模型为ViT-L。



通过在ImageNet-1K（224尺寸）上进行微调评估MAE预训练与有监督预训练的对比。

与在IN1K或JFT300M上对训练的原始ViT结果比较。



ViT-L在部分微调的Transformer 块fine-tuning 的结果。

调 0 个块是线性探测；

24 是完整的 fine-tuning。

MAE 表示的线性可分性较差，但当一个或多个块被调优时，它始终优于 MoCo v3。

**Conclusion:**

简单且能很好地扩展的算法是深度学习的核心。在NLP中，简单的自监督学习方法可以使模型规模扩展时获得更多好处。在计算机视觉中，尽管自监督学习有所进步，但实用的预训练范例主要是监督式的。在ImageNet上和在迁移学习中，自编码器—一种与NLP技术相似的简单自监督方法—提供了可扩展的好处。视觉中的自监督学习现在可能正沿着类似于NLP的轨迹发展。

图像和语言是不同性质的信号。图像只是记录的光线，没有分解成视觉上的“单词”。MAE不是尝试移除物体，而是移除很可能不构成语义部分的随机patch。同样，MAE重构不是语义实体像素。然而，MAE能推断出复杂的整体重构，这表明它已学习到许多视觉概念，即语义。它通过MAE内部的丰富隐藏表示实现。

更广泛的影响：基于训练数据集的学习统计信息来预测内容，会反映这些数据的偏向，包括具有负面社会影响的偏向。可能会生成不存在的内容。这些问题需要进一步研究和考虑。

**Related Work：**

Masked language modeling：

BERT and GPT : These methods hold out a portion of the input sequence and train models to predict the missing content.

Autoencoding:

It has an encoder that maps an input to a latent representation and a decoder that reconstructs the input.

Works:PCA,DAE

Masked image encoding:

Masked image encoding methods learn representations from images corrupted by masking

Pioneering work:DAE

Works: iGPT,ViT,BEiT

Self-supervised learning:

Often focusing on different pre-text tasks for pre-training.

Self-supervised learning is a type of machine learning that involves training a model using unlabeled data and a small amount of labeled data. The goal is to extract information from the unlabeled data and use it to perform a task for which labeled data is available.

**Approach:**

Masking：

we divide an image into regular non-overlapping patches. Then we sample a subset of patches and mask the remaining ones. we sample random patches without replacement, following a uniform distribution.

MAE encoder：

Encoder is a ViT, a series of Transformer blocks.

MAE decoder:

The input to the MAE decoder is the full set of tokens consisting of (i) encoded visible patches, and (ii) mask tokens.

The MAE decoder is only used during pre-training to perform the image reconstruction task (only the encoder is used to produce image representations for recognition)

Reconstruction target:

Our MAE reconstructs the input by predicting the pixel values for each masked patch.

Simple implementation:

First we generate a token for every input patch (by linear projection with an added positional embedding). Next we randomly shuffle the list of tokens and remove the last portion of the list, based on the masking ratio. This process produces a small subset of tokens for the encoder and is equivalent to sampling patches without replacement. After encoding, we append a list of mask tokens to the list of encoded patches, and unshuffle this full list (inverting the random shuffle operation) to align all tokens with their targets

**QUESTIONS:**

-what did authors try to accomplish?

A MAE approach that first accomplicated in NLP fields to improves training speeds for computer vision deep models.

-what the key elements of the approach?

An encoder-decoder arthitecture.

-what can you use yourself?

This MAE approach and new understanding of encoder-decoder arthitecture.

-what other reference do you want to follow?

GPT and BERT.