Masked Autoencoders Are Scalable Vision Learners

CVPR 2022

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár and Ross Girshick

Fackbook AI Research (FAIR)

2022 年 12 月 12 日

Overview

目的

- 以一定比例随机 mask 掉图片中的一些图像块 (patch) 然后重建这些部分的像素值

核心设计

- 非对称的编码-解码结构
- 屏蔽掉 75% 的像素再重构这些像素很有意义

优点

- 速度快：训练时间缩短三倍或以上
- 精度高：在 ImageNet-1K 上达到了 87.8% 的准确率

Motivation

- 现有模型对数据量要求越来越大，但没有这么多 labeled 图像
- NLP 通过自监督的预训练解决了对 label 的依赖 (NLP 中标签数据很少)
  1. GPT (ChatGPT 前前前身): 先在大规模语料上进行无监督预训练, 再在小得多的有监督数据集上为具体任务微调 (fine-tune)
  2. BERT: 提出 masked language model（MLM）减少对 label 的依赖，然后进行具体任务微调

GPT & BERT

GPT

- Generative Pre-Training - 2018 OenAI
- 无监督 Pre-training 和有监督 Fine-tuning 目的：学习一种通用的 Representation 方法，针对不同种类的任务只需略作修改便能适应
- GPT 使用句子序列预测下一个单词，因此要采用 Mask Multi-Head Attention 对单词的下文遮挡，防止信息泄露

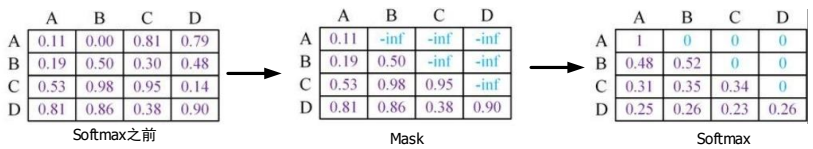

图 1: 经过 Mask 和 Softmax 之后，当 GPT 根据单词 A 预测单词 B 时，只能使用单词 A 的信息，根据 [A, B] 预测单词 C 时只能使用单词 A, B 的信息

GPT & BERT

BERT

- Pre-training of Deep Bidirectional Transformers for Language Understanding - 2019 Google
- Masked Language Model (MLM): 15% 的概率选中某个 token，按照以下策略 mask
  1. 80% 直接 mask 掉, I am a good **man** → I am a good **[MASK]**
  2. 10% 替换成其他库中的 token, I am a good **man** → I am a good **woman**
  3. 10% 替换成本句中其他 token, I am a good **man** → I am a good **I**
- 这样学到的模型对 [MASK] 和所有 token 都敏感

GPT & BERT

BERT

图 2: Bert 与芝麻街 (Sesame Street) 中人物同名，因此介绍 Bert 时会用到动漫图片，现在 Google 的芝麻街系列已经有 6 个成员了

Mask 思想在 CV 和 NLP 中的不同

- 框架不同：在以卷积为基础的 CV 中，没有位置编码和 token 这些概念。但 ViT 解决了
- 信息密度不同：语言信息密度极高，图像偏低，mask 后甚至可以插值重构。通过高比例mask 掉patch解决了
- 自编码器中的 decoder：CV 中，decoder 重构的是低语义的像素；NLP 中重构的是高语义的缺失单词

Overview

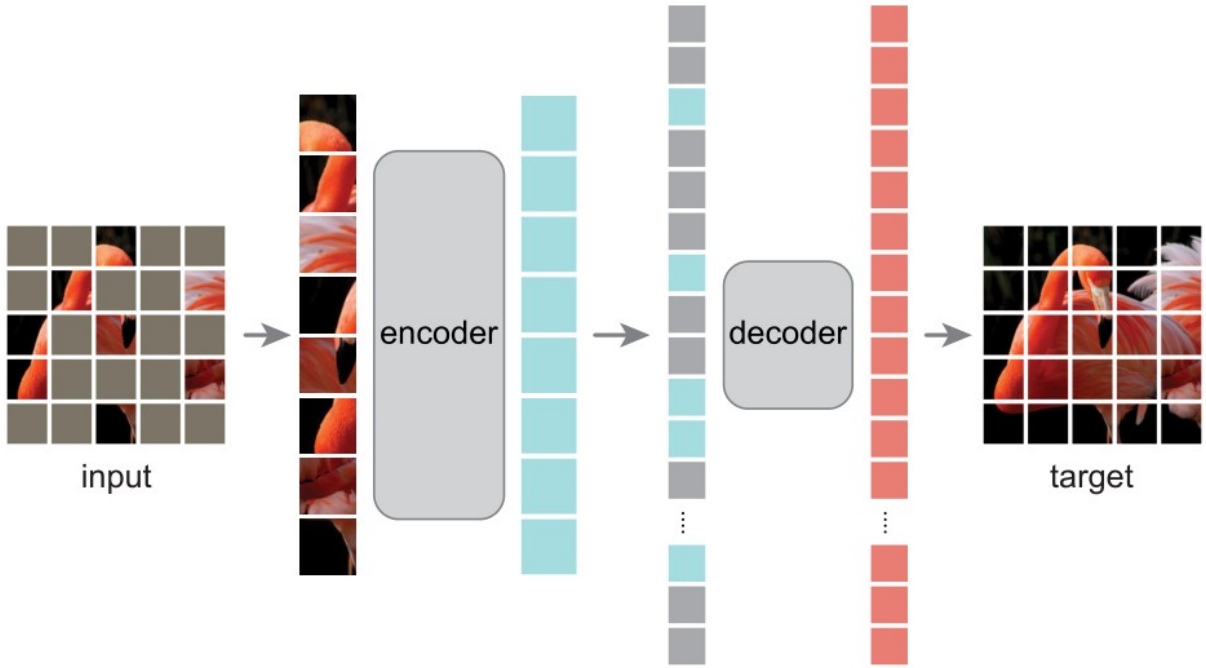

图 3: Masked Autoencoder 流程图。包括一个非对称的编码-解码结构，编码器只关心 unmasked 区域 (节省时间并提升精度)，而解码器根据位置编码和所有的 patches 来重构图像

Masking

- 将图像划分为规则的不重叠的 patches
- 随机选取 mask 区域
- 这些抽样遵循均匀分布

MAE encoder

- 通过添加位置编码的线性投影嵌入 patches，然后通过一系列 Transformer 块处理结果集合
- 占比很高的 Masked patches 被直接忽略，因此减少了内存和计算量

MAE decoder

- 处理的是所有的 patches
- 每个 masked patch 是一个可学习的向量，目的就是优化这些向量，使得对应的 patch 尽可能恢复到 mask 前的样子
- 完整的位置编码用来规定 patches 对应的位置
- 这里设计的 decoder 只是用在预训练阶段，用在具体任务中可以设计其它形式的 decoder

Implementation

- 如何约束模型：在这些masked patches 上用均方误差MSE做损失函数
- 实验中约束的是归一化的像素值，提高重构质量

实验设置

实验方式

- Baseline: ViT-Large (这个模型很大，很容易过拟合)
- 训练策略：在 ImageNet-1K 100 万张图片上自监督预训练，再在同样的数据集上做有标号的监督训练
- 做法
  1. 端到端微调 (end-to-end fine-tuning)
  2. 线性探测 (linear probing, 将最后一层替换为线性层进行分类，只训练这个层，比较效果)
- 结果：top-1 的精度

消融实验

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

(a) **Decoder depth**. A deep decoder can improve linear probing accuracy.

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

(b) **Decoder width**. The decoder can be narrower than the encoder (1024-d).

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

(c) **Mask token**. An encoder without mask tokens is more accurate and faster (Table 2).

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

(d) **Reconstruction target**. Pixels as reconstruction targets are effective.

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

(e) **Data augmentation**. Our MAE works with minimal or no augmentation.

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

(f) **Mask sampling**. Random sampling works the best. See Figure 6 for visualizations.

图 4: MAE 和 ViT-16 在 ImageNet-1K 上的消融实验; 默认配置用 灰色 标出; Depth 指 Transformer 块的数目; Width 指每个 token 表示成多长的向量; Mask token 指要不要在 encoder 中加入 masked patch; Reconstruction target 指最终的优化目标对实验的影响, dVAE token 是 BEiT 的做法, 通过 ViT 把每一个 patch 映射到一个离散的 token, 像 BERT 一样的去做预测; Data augmentation 指裁剪方式的影响; Mask sampling 指以什么样的方式 mask patches(后面有说明)

消融实验

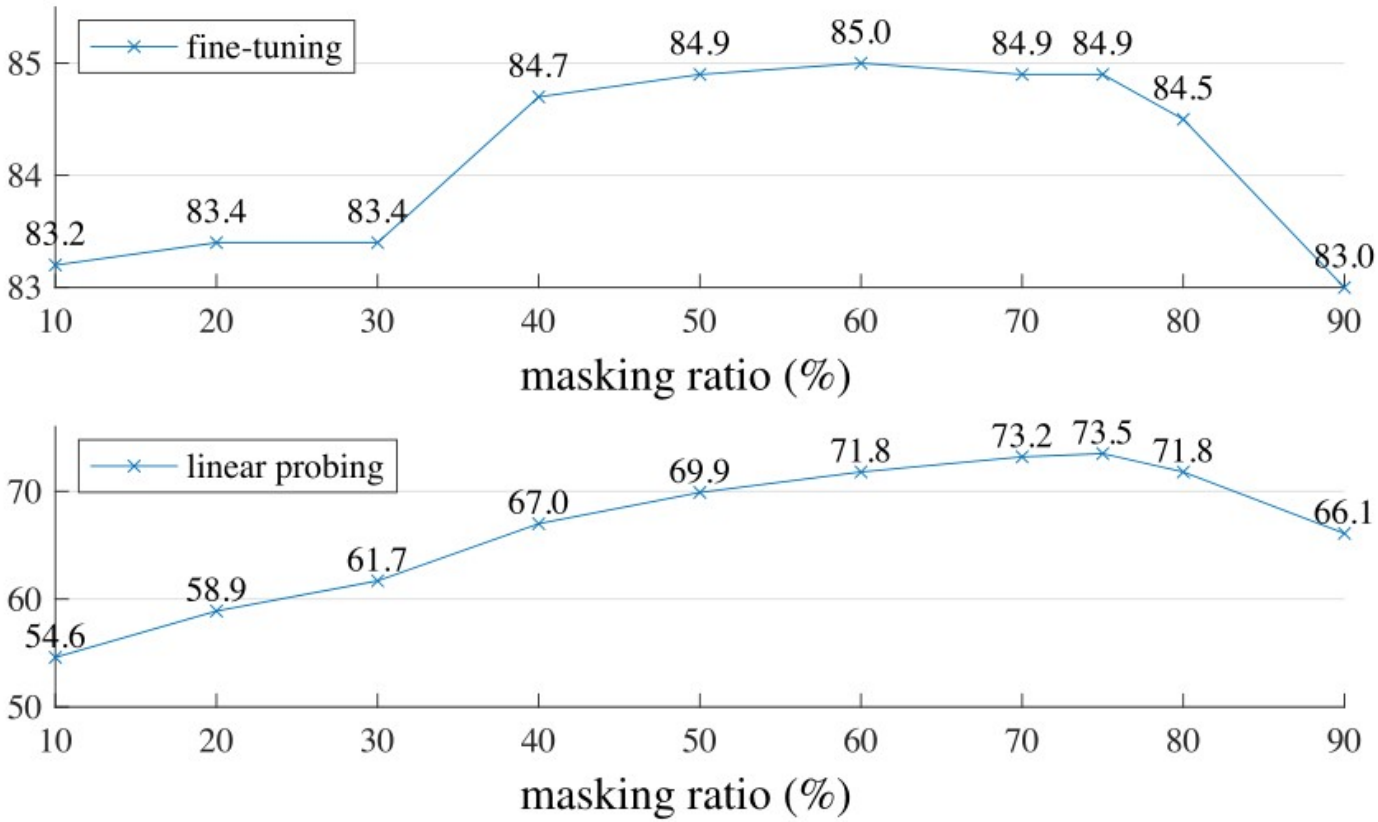

图 5: Masking 比例对实验结果的影响，最终选择了 75%

消融实验

encoder	dec. depth	ft acc	hours	speedup
ViT-L, w/ [M]	8	84.2	42.4	-
ViT-L	8	84.9	15.4	2.8×
ViT-L	1	84.8	11.6	**3.7**×
ViT-H, w/ [M]	8	-	119.6†	-
ViT-H	8	85.8	34.5	3.5×
ViT-H	1	85.9	29.3	**4.1**×

图 6: 训练时间，加速效果很明显

消融实验

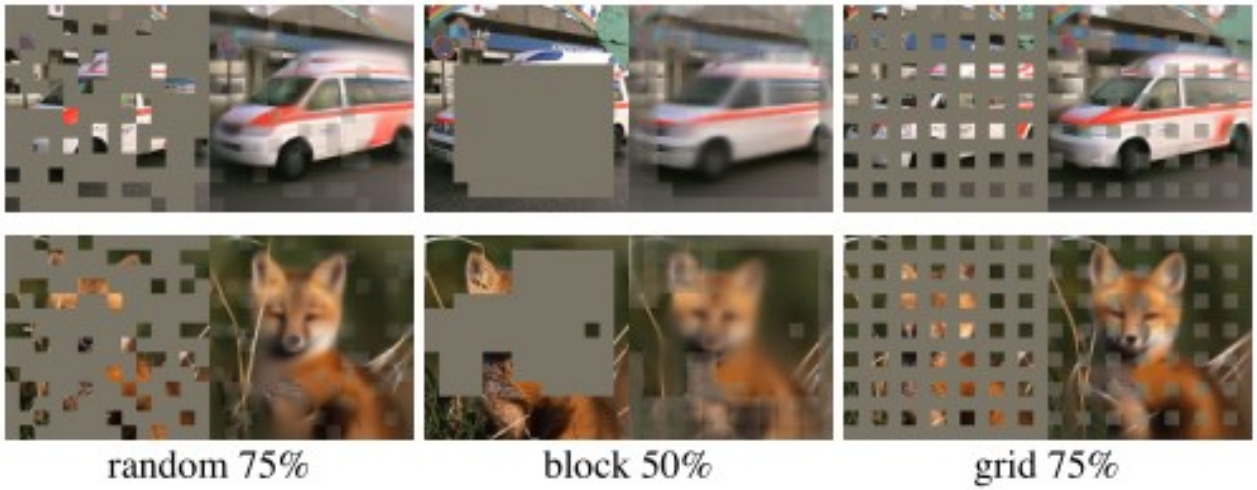

图 7: 不同 mask 方式的示意图，随机均匀的效果更好

对比实验

method	pre-train data	ViT-B	ViT-L	ViT-H	ViT-H_{448}
scratch, our impl.	-	82.3	82.6	83.1	-
DINO [5]	IN1K	82.8	-	-	-
MoCo v3 [9]	IN1K	83.2	84.1	-	-
BEiT [2]	IN1K+DALLE	83.2	85.2	-	-
MAE	IN1K	83.6	85.9	86.9	**87.8**

图 8: 与基于 ViT 的自监督方法对比，可以发现纵向差距不是很大，而横向差距明显，说明主要挑战是过拟合问题

迁移学习

应用到下游任务中

- 物体检测和分割
- 语义分割
- 分类
- 重建 (eg. 像素，符号)

写作

- 朴实的标题: XX is a good (scalable, efficient) XX
- Organization 清晰: 多用问句引出

实验

- 由 masked patches 想到可以在训练数据中增加噪声来减少过拟合
- 在无人机定位中，卫星图像是不是可以编码成较高级的表示，每次只需要编码无人机图像，比较两个向量 (猜测)

结果是不是太好了

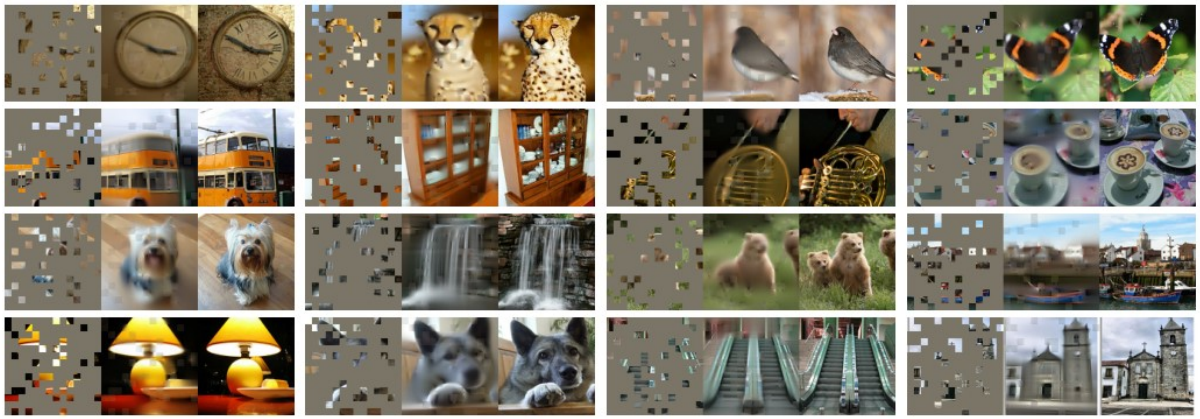

Figure 2. Example results on ImageNet *validation* images. For each triplet, we show the masked image (left), our MAE reconstruction† (middle), and the ground-truth (right). The masking ratio is 80%, leaving only 39 out of 196 patches. More examples are in the appendix.
†*As no loss is computed on visible patches, the model output on visible patches is qualitatively worse. One can simply overlay the output with the visible patches to improve visual quality. We intentionally opt not to do this, so we can more comprehensively demonstrate the method's behavior.*

Figure 3. Example results on COCO validation images, using an MAE trained on ImageNet (the same model weights as in Figure 2). Observe the reconstructions on the two right-most examples, which, although different from the ground truth, are semantically plausible.

图 9: 在 ImageNet 和 COCO 上测试的一些例子。预训练阶段的数据对结果的影响还是很大的，有些图像根本不可能重构出来，可能因为训练时学到了相似的偏置

Thanks!