Masked Autoencoders Are Scalable Vision Learners

2021250020 한정찬

01

Introduction

Masked autoencoder

NLP

Autoregressive modeling of GPT

Masked autoencoding of BERT

➔ Remove a portion of data, learn to predict

Computer Vision

Masked autoencoder
➔ don't work well

What's the difference ??

Language vs Image

- Architectures
  - ➔ Attention VS CNN (solved by ViT)
- Information density
  - ➔ Language > Image (하나의 sematic 정보
  - ➔ Spatial redundancy
- Decoder's role
  - ➔ predict words vs. pixels

02

Methodology

Our MAE

- High Masking ratio
  - ➔ reduce redundancy
- Asymmetric encoder-decoder
  - ➔ Encoder – only patches without masking
  - ➔ Decoder – latent representation + masked
- Reduce computation

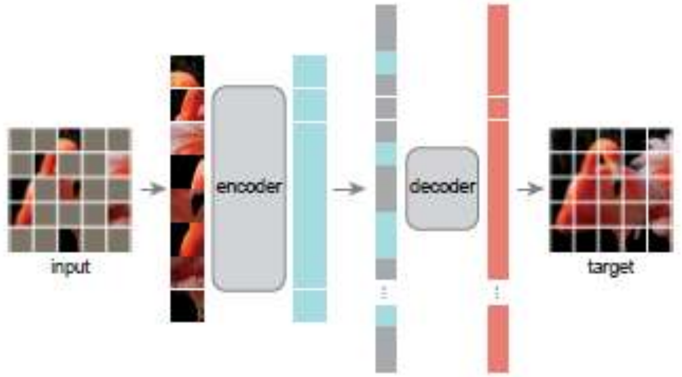


Figure 1. **Our MAE architecture**. During pre-training, a large random subset of image patches (*e.g.*, 75%) is masked out. The encoder is applied to the small subset of *visible patches*. Mask tokens are introduced *after* the encoder, and the full set of encoded patches and mask tokens is processed by a small decoder that reconstructs the original image in pixels. After pre-training, the decoder is discarded and the encoder is applied to uncorrupted images (full sets of patches) for recognition tasks.

Our MAE

Masking

- Divide image into non-overlapping patches
- Random sampling with high masking ratio

MAE encoder

- ViT but applied only on unmasked patches (25%)
- Less computation, memory

Our MAE

MAE decoder

- Input: encoded patches + masked tokens
- Add positional embeddings (info about location)
- MAE decoder is only used during pre-training reconstruction task
- Independent of encoder design

Reconstruction

- Predict pixel values of masked patches
- Loss function computes MSE only on masked patches(like BERT)

03

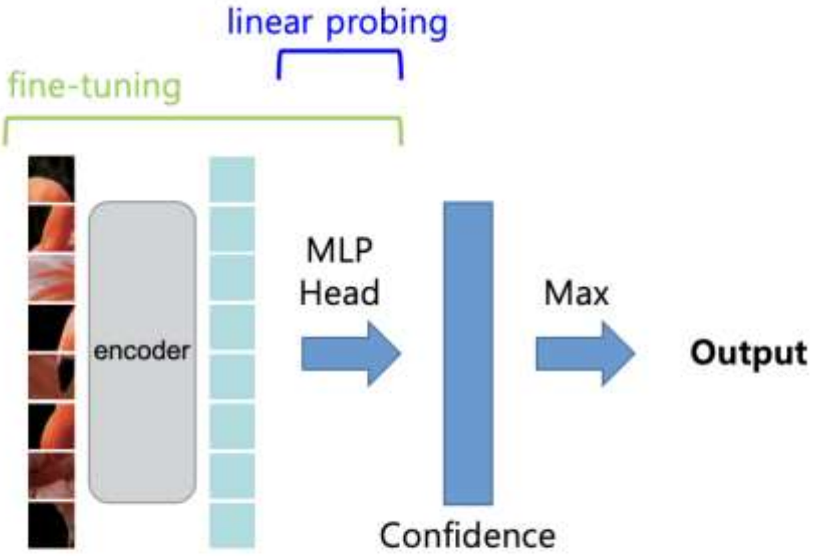
Experiments

ImageNet Experiments

- Self-supervised pre-training on ImageNet-1K

- Then evaluate
  ① Fine-tuning
  ② Linear probing

- Baseline: ViT-Large



Masking ratio

- High Masking ratio

  ➔75% is good for both

- Contrast behavior with BERT (typically 15%)

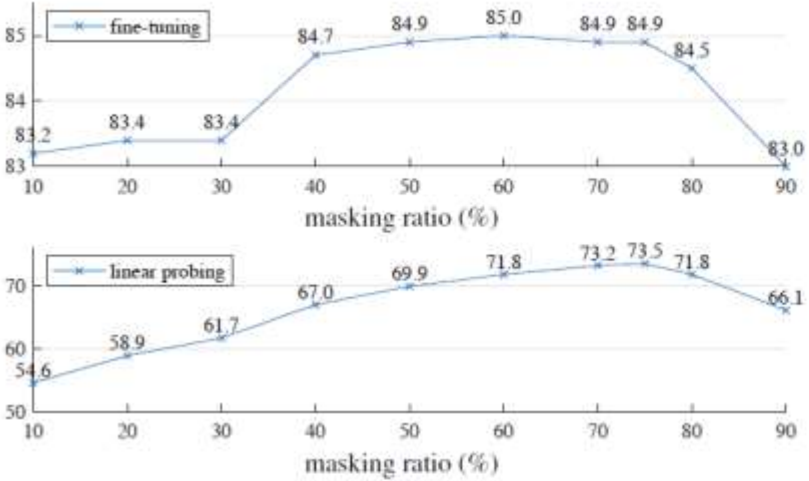


Figure 5. **Masking ratio**. A high masking ratio (75%) works well for both fine-tuning (top) and linear probing (bottom). The y-axes are ImageNet-1K validation accuracy (%) in all plots in this paper.

Decoder Design

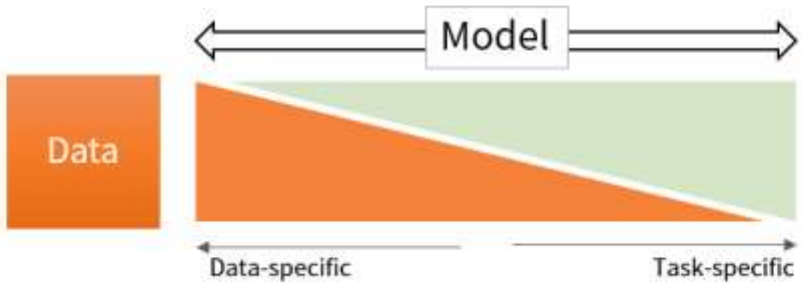
blocks	ft	lin
1	84.8	65.5
2	**84.9**	70.0
4	**84.9**	71.9
8	**84.9**	**73.5**
12	84.4	73.3

(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).

- sufficiently deep decoder is important for linear probing
- gap between a pixel reconstruction task and a recognition task



Others

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

- Reduce computation

(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.

Comparisons with previous results

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**

Conclusion

- Language -> BERT
- Vision -> MAE

- Different domain, same approach, meaningful result

Thanks!