Masked Autoencoders Are Scalable Vision Learners

Kaiming He (Facebook AI Research), 2021

Chioh Lee
202555439
Dept. of AI, School of CSE

https://arxiv.org/abs/2111.06377

Abstract, Introduction (Crash Course)

Rationales:

(1) ViT = Data-hungry

(2) GPT: Autoregressive Modeling

(3) BERT: Masked Autoencoding

(4) Image has heavy spatial redundancy

(mask can be reversed w/ extrapolation)

Abstract, Introduction (Crash Course)

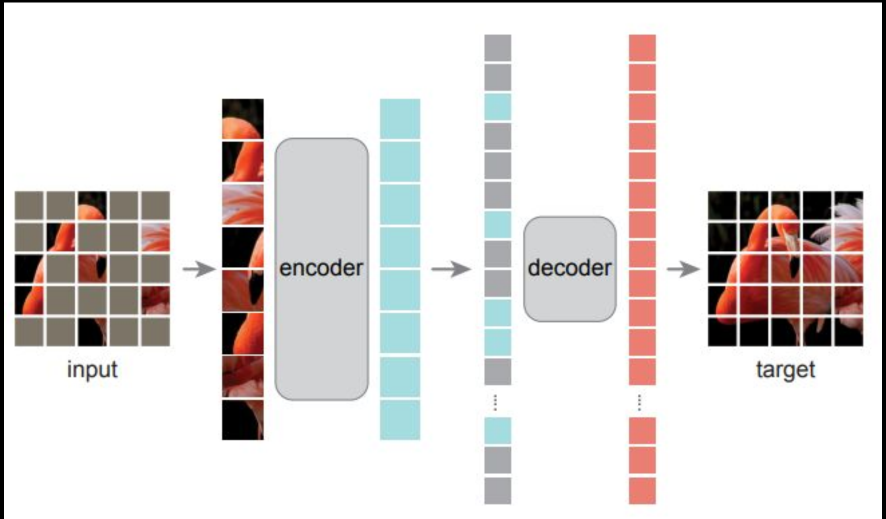
Methodologies:

(1) Mask image patches

→ Reconstruct

(2) Asymmetric Encoder-Decoder

(smaller decoder)



Findings:

(1) Masking significant portion of image

→ nontrivial performance increase

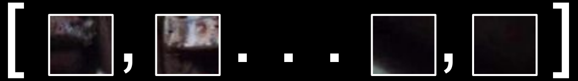
(2) Training Acceleration (> 3x)

Approach: Patchfying

$$x \in \mathbb{R}^{H \times W \times 3}$$

$$P \times P$$

$$N = \frac{H \cdot W}{P^2}$$



default D = 1024

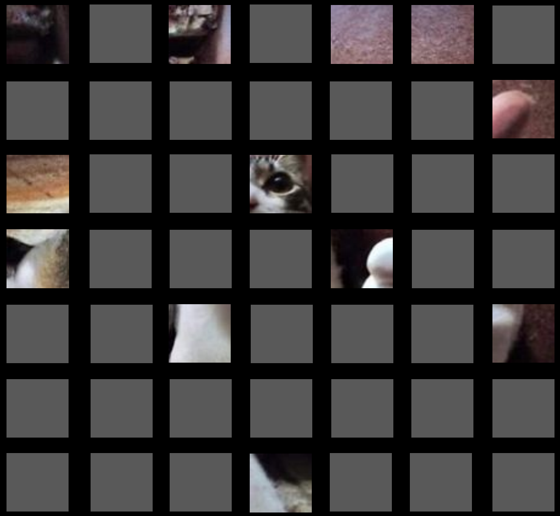
$$z_i = \mathrm{PatchEmbed}(x_i), \quad z_i \in \mathbb{R}^D \quad \text{for } i = 1, \ldots, N$$

$$Z = [z_1, z_2, \ldots, z_N] \in \mathbb{R}^{N \times D}$$

Approach: Masking ← rationale (4)

Z

(I know, just assume this for easy visualization)



$$V \subset \{1, \ldots, N\}, \quad |V| = \lfloor r \cdot N \rfloor$$

$$V \sim \mathrm{Uniform}(\{1, \ldots, N\})$$

[ . . .] + PE

$$\tilde{Z}_V = \{z_i + p_i \mid i \in V\}$$

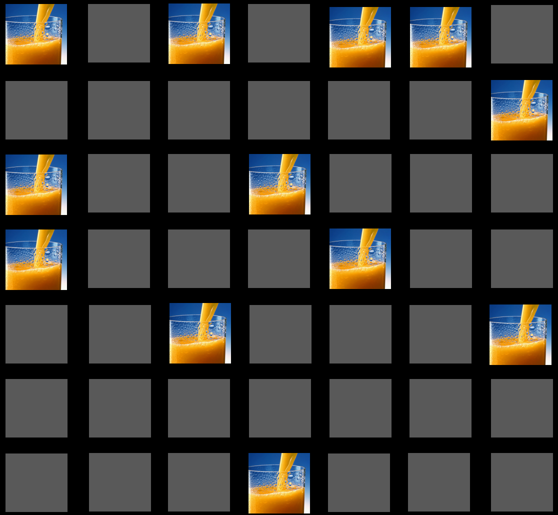
Approach: Encoder

$$\tilde{Z}_V$$

(Again, assume)

$$h_V = f_{\text{enc}}(\tilde{Z}_V)$$

Approach: Decoder

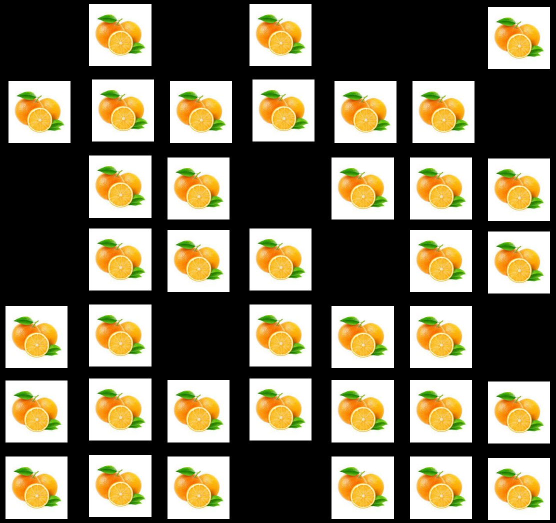


$$z_{\text{mask}} \in \mathbb{R}^D$$

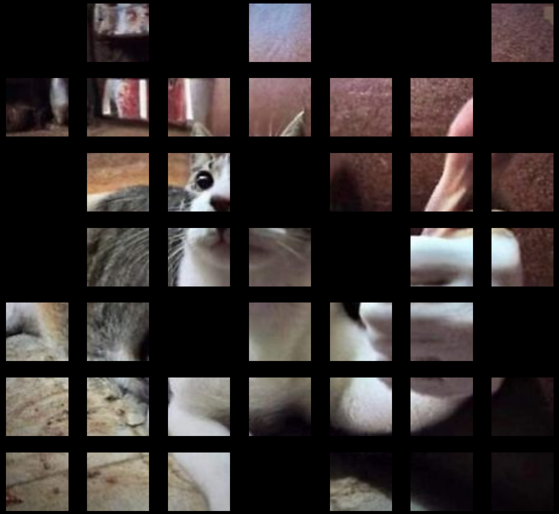
$$\tilde{Z}_{\text{dec}} = \text{Sort}(\{h_i\}_{i \in V} \cup \{z_{\text{mask}} + p_i\}_{i \in M})$$

$$\hat{x}_i = f_{\text{dec}}(\tilde{Z}_{\text{dec}})_i \quad \text{for } i \in M$$

Approach: Reconstruction Target

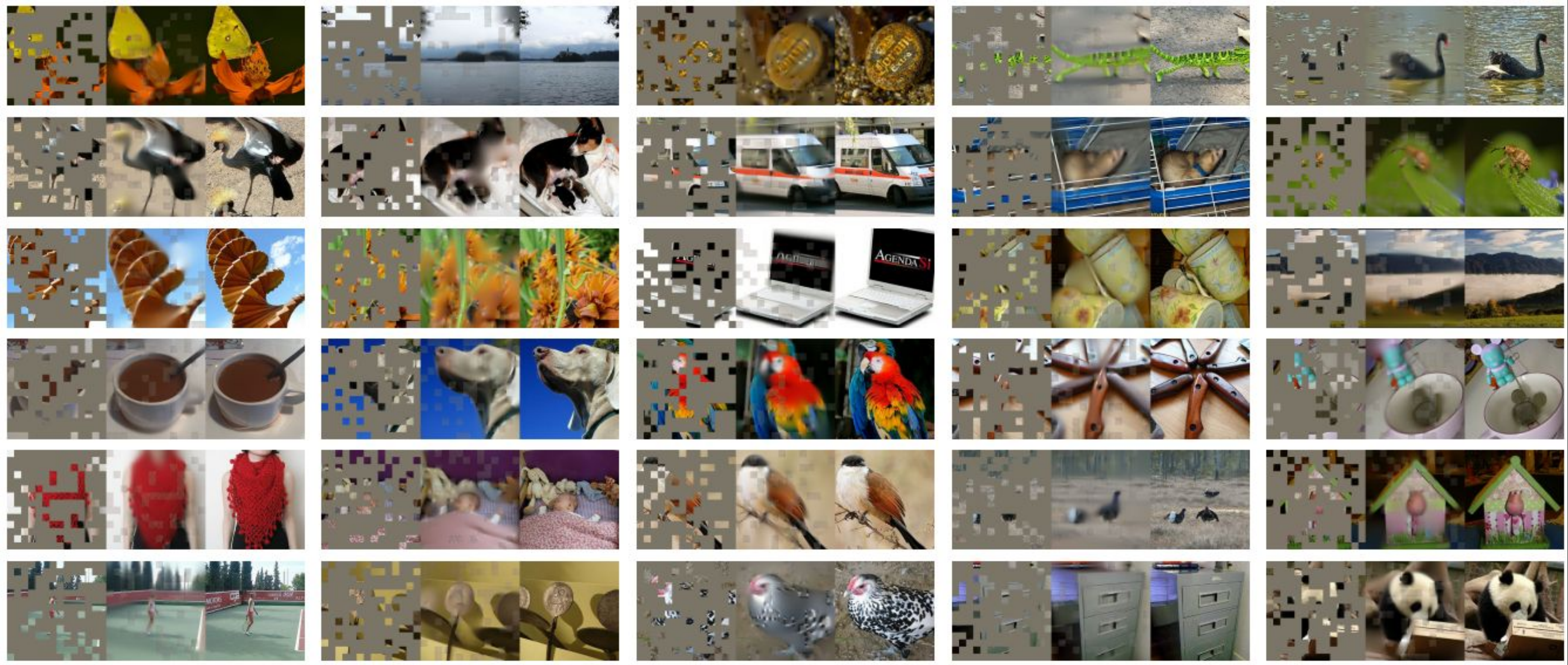


(Normalization)

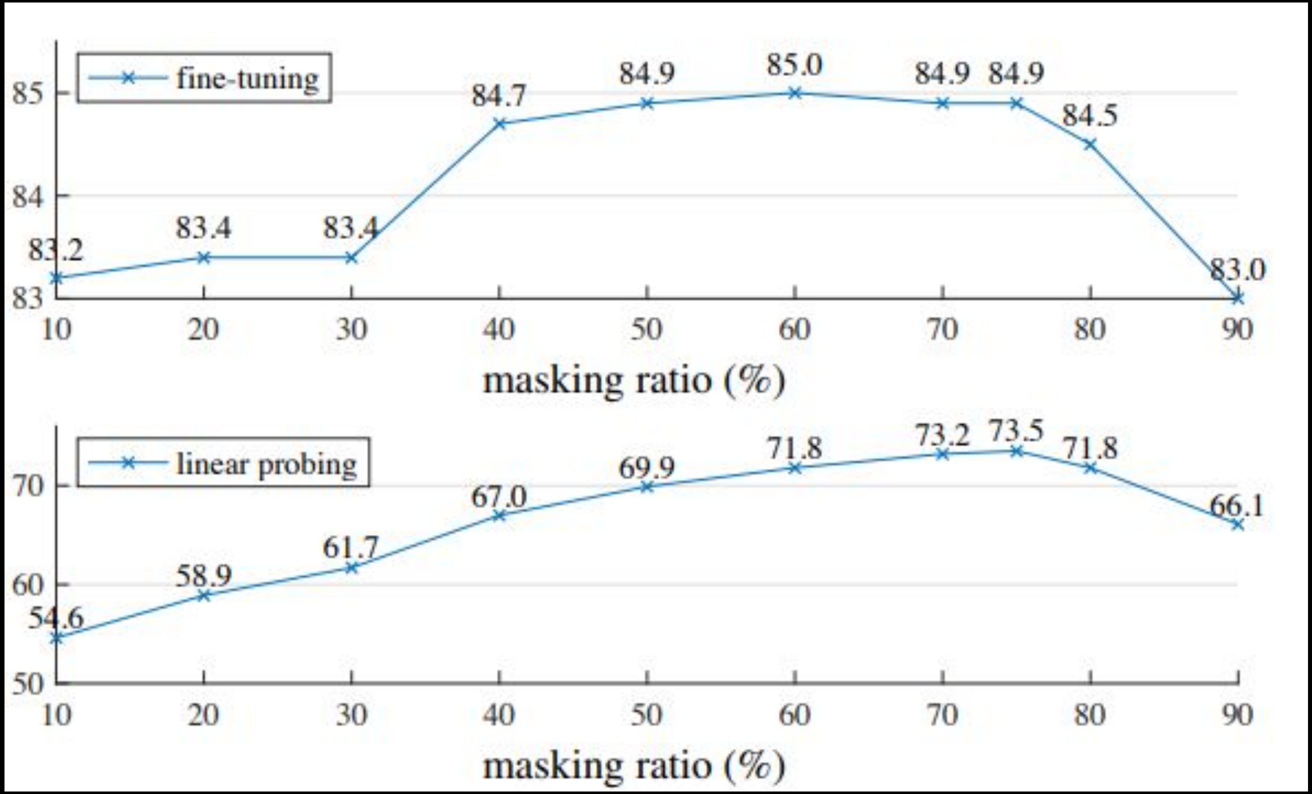


$$\mathcal{L} = \frac{1}{|M|} \sum_{i \in M} \|\hat{x}_i - x_i\|^2$$

Approach: Results



Experiments: Masking Ratio



Experiments: Depth, Accuracy, Train

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**×

Experiments: Ablation Studies

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.

Questions

Q. Use of smaller decoder?

In vision, the decoder reconstructs pixels, hence its output is of a lower semantic level than common recognition tasks; different from BERT, where the decoder predicts missing words that contain rich semantic information.

Q. Use of mask tokens in decoder input?

If the encoder uses mask tokens, it performs worse. The encoder has a large portion of mask tokens in its input in pretraining, which does not exist in uncorrupted images. By skipping the mask token in the encoder, this may also reduce training computation.