Masked Autoencoders Are Scalable Vision Learners

{Kaiming He; Xinlei Chen Saining Xie Yanghao Li Piotr Doll´ar Ross Girshick} @ Facebook AI Research

https://arxiv.org/abs/2111.06377

Hyunseok Lim

2024.05.13

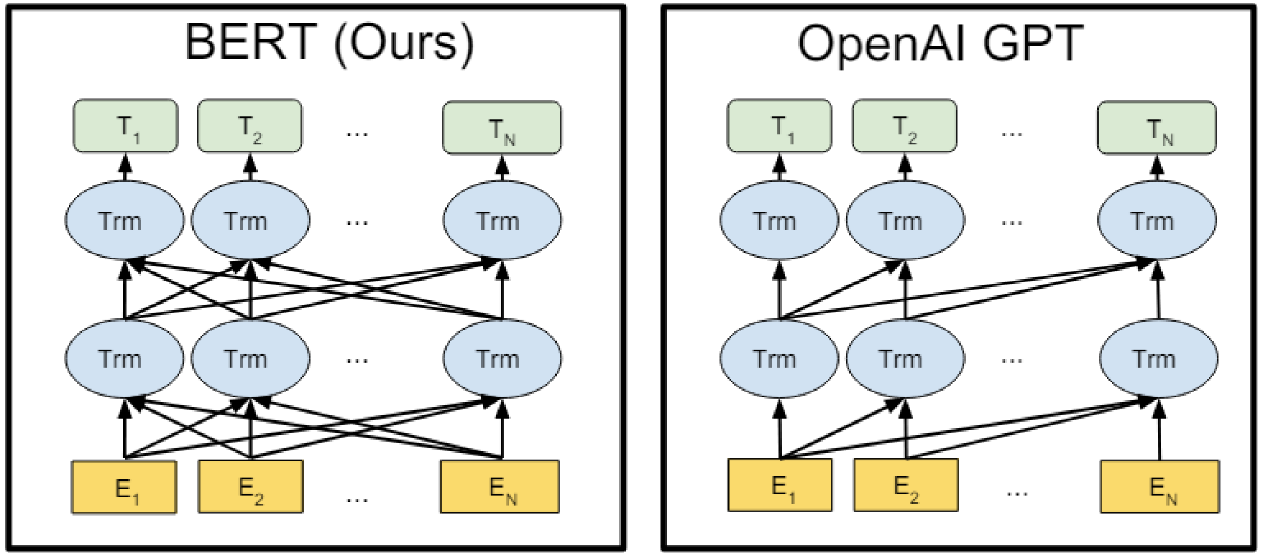
Introduction: Motivation

Deep learning models today can easily overfit one million images and begin to demand hundreds of millions often publicly inaccessible—labeled images.

Scale-up model size ▶ Requires more data ▶ Difficulty getting labeled data

This appetite for data has been successfully addressed in natural language processing (NLP) by ***self-supervised pretraining***.

They *remove a portion of the data* and *learn to predict the removed content*. (Pretext task ➔ 데이터 자체에 대한 이해도 ↑)



Masked autoencoding

Auto-regressive

We ask:

what makes masked autoencoding different between vision and language?

Introduction: Different between CV and NLP

1. Convolution and Transformer

Conv kernel처럼 규칙적인 그리드에서 작동하며,
Mask token이나 positional embedding을 Convolution에 통합하는 것은 쉽지 않음
Vision Transformers (ViT)의 등장으로 제약이 많이 해소됨

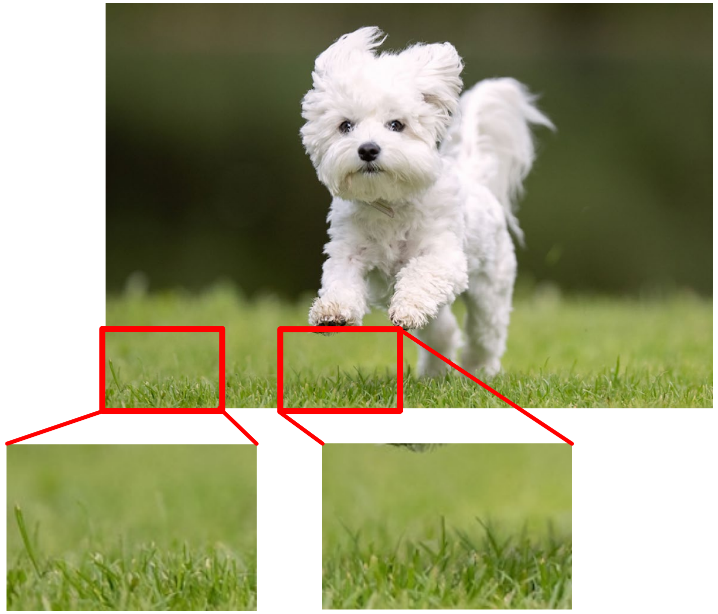
2. Information Density

Highly semantic

Information-dense

"Information density is different between language and vision."

Sophisticated language understanding



heavy spatial redundancy

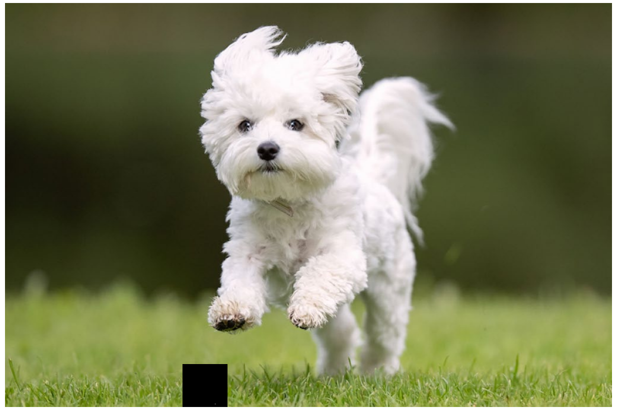
3. Decoder plays a different role

NLP's decoder predicts missing words that contain rich semantic information

"Information density is different between language and vision."

Rich semantic information

CV's decoder reconstructs pixels that has lower semantic level



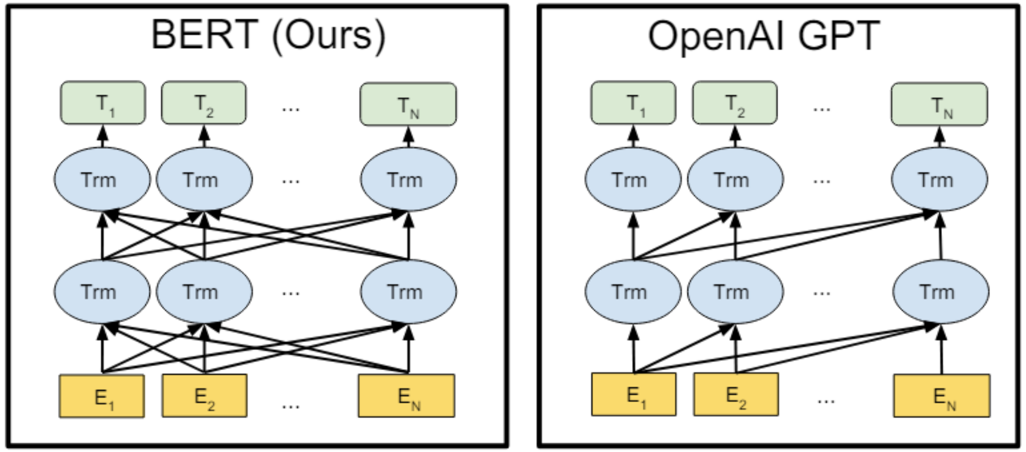
 Low semantic information

Related works

Masked language modeling

- BERT
- GPT

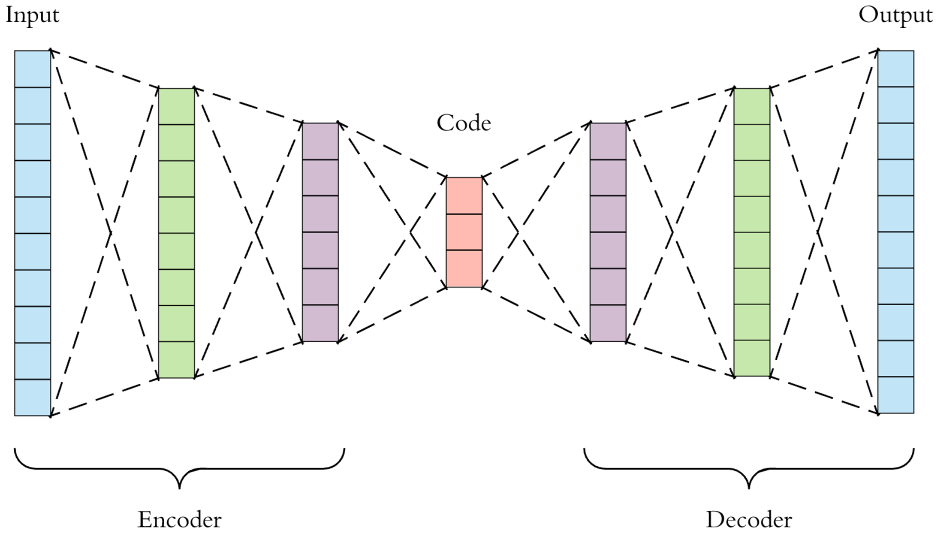
- these pre-trained representations generalize well to various downstream tasks
- **많은 데이터를 학습할 수록** emergent ability **향상**



Autoencoding

- Autoencoder
- Denoising Autoencoder (DAE)

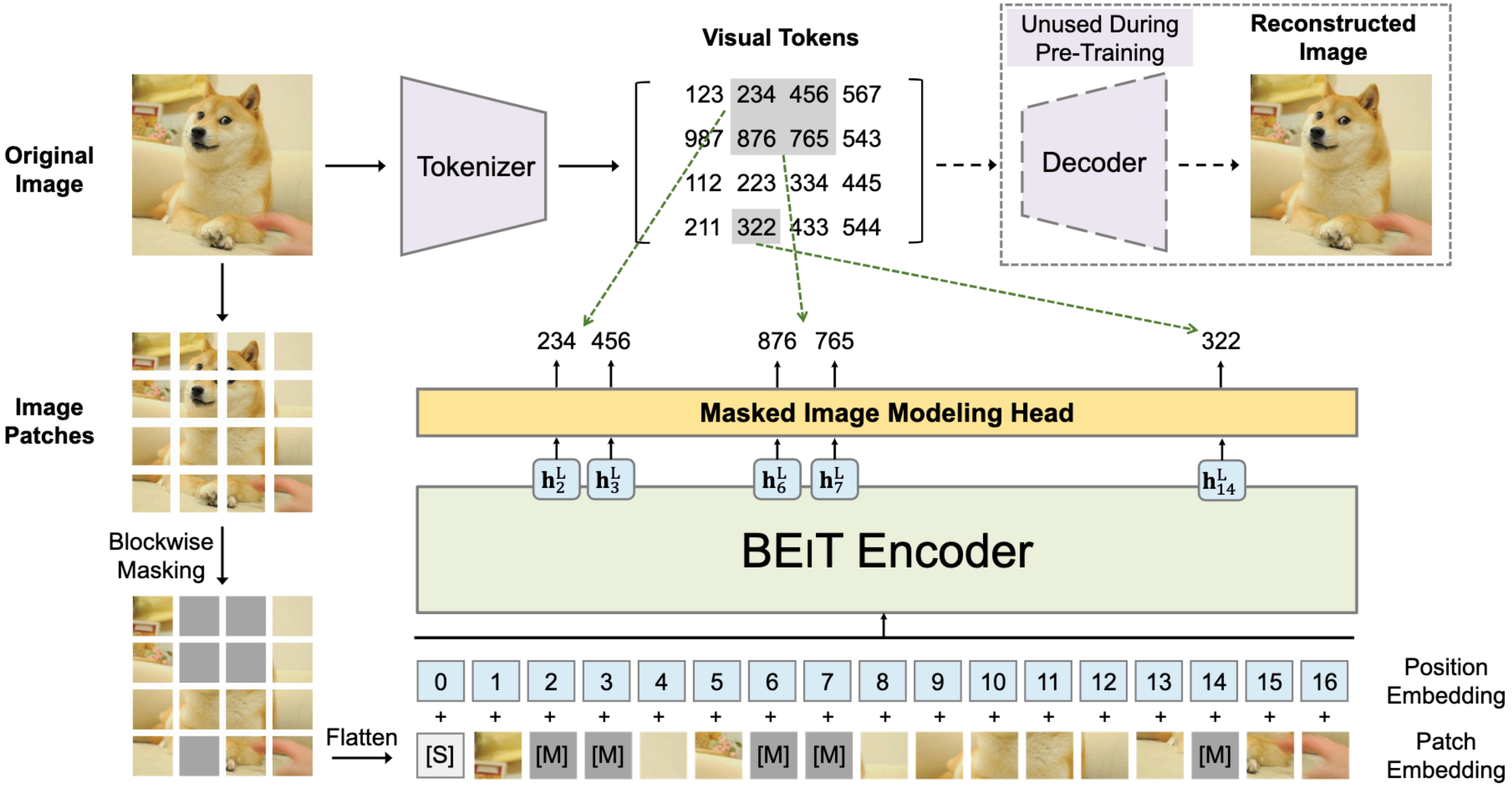
Masked AE**는** DAE**의 한 종류이지만**, Classical AE**와는 구조적으로 다름**



Masked Image Encoding

- Context Encoder inpaints large missing regions using convolutional networks
- BEiT proposes to predict discrete tokens

Context encoding

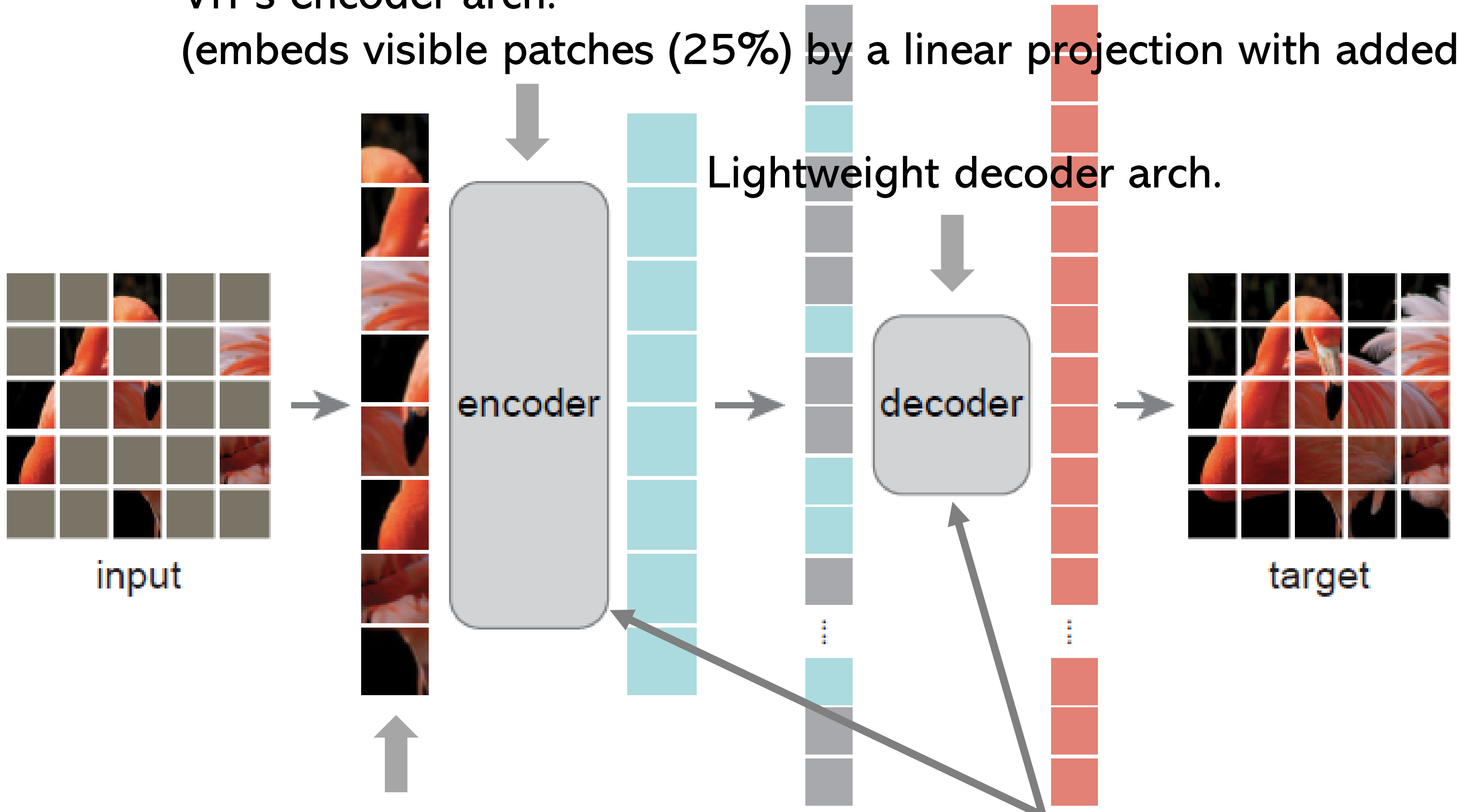


BEiT: BERT Pre-training of Image Transformers

Approach: Architecture

ViT's encoder arch.
(embeds visible patches (25%) by a linear projection with added positional embeddings)

Lightweight decoder arch.

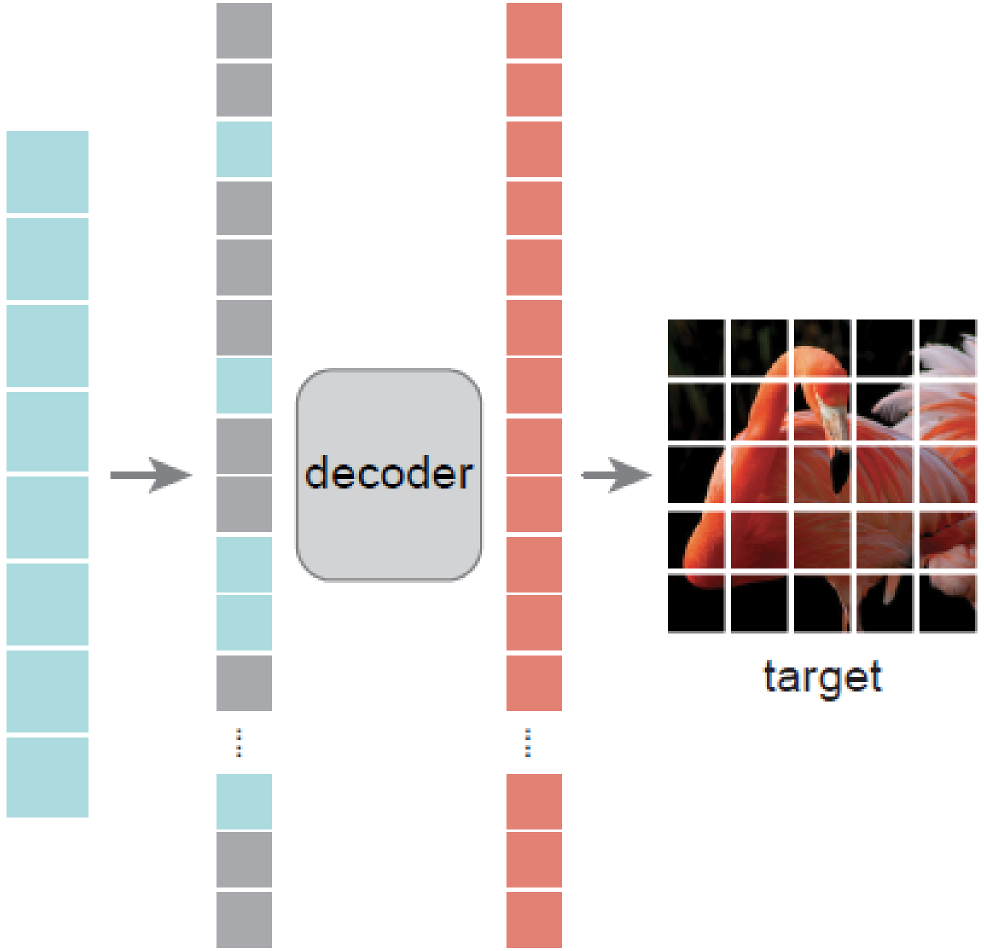


Sampling random patches (with high masking ratio),
following a <u>uniform</u> distribution

Asymmetric encoder-decoder design

Masking ratio(75%) 를 높이면 주변 패치들의 정보를 참고하여 쉽게 예측하지 못하게 함

Approach: Architecture



Visible patch + Masked patch
+ Positional embedding → Decoder

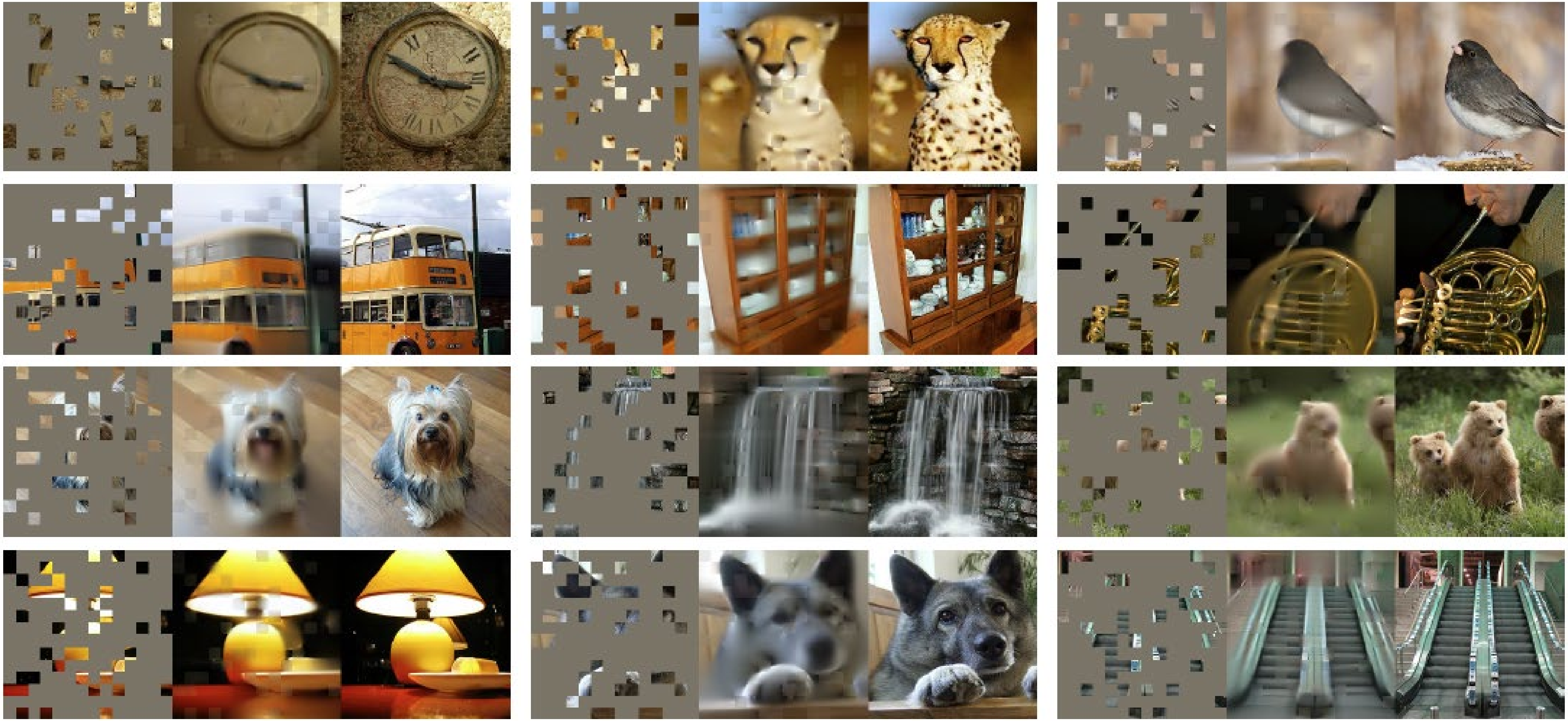
Decoder는 Pretraining (Reconstruction task)단계에서만 사용됨

Encoder의 Arch.와 독립적으로 유연하게 설계할 수 있음

Encoder 대비 10% 정도의 Computation cost가 요구됨

모델의 학습 속도 향상
(This can reduce overall pre-training time by 3x or more)

Approach: Architecture



Approach: Architecture

Reconstruction target

Loss function computes the mean squared error (MSE) between the reconstructed and original images

We compute the loss only on masked patches, similar to BERT. (differs from traditional denoising autoencoders)

This choice is purely result-drivThis choice is purely result-driven: computing the loss on all pixels leads to a slight decrease in accuracy (e.g., 0.5%).

en: computing the loss on all pixels leads to a slight decrease in accuracy (e.g., 0.5%).

ImageNet Experiments

Self-supervised pre-training on the ImageNet 1K Dataset(IN1K) training set

Supervised training to evaluate the representations with **end-to-end fine-tuning** or **linear probing**

Baseline: ViT-Large is used as backbone in ablation study
ViT-L is very big and tends to overfit
Strong regularization is needed

scratch, original [16]	scratch, our impl.	baseline MAE
76.5	82.5	84.9

ImageNet Experiments

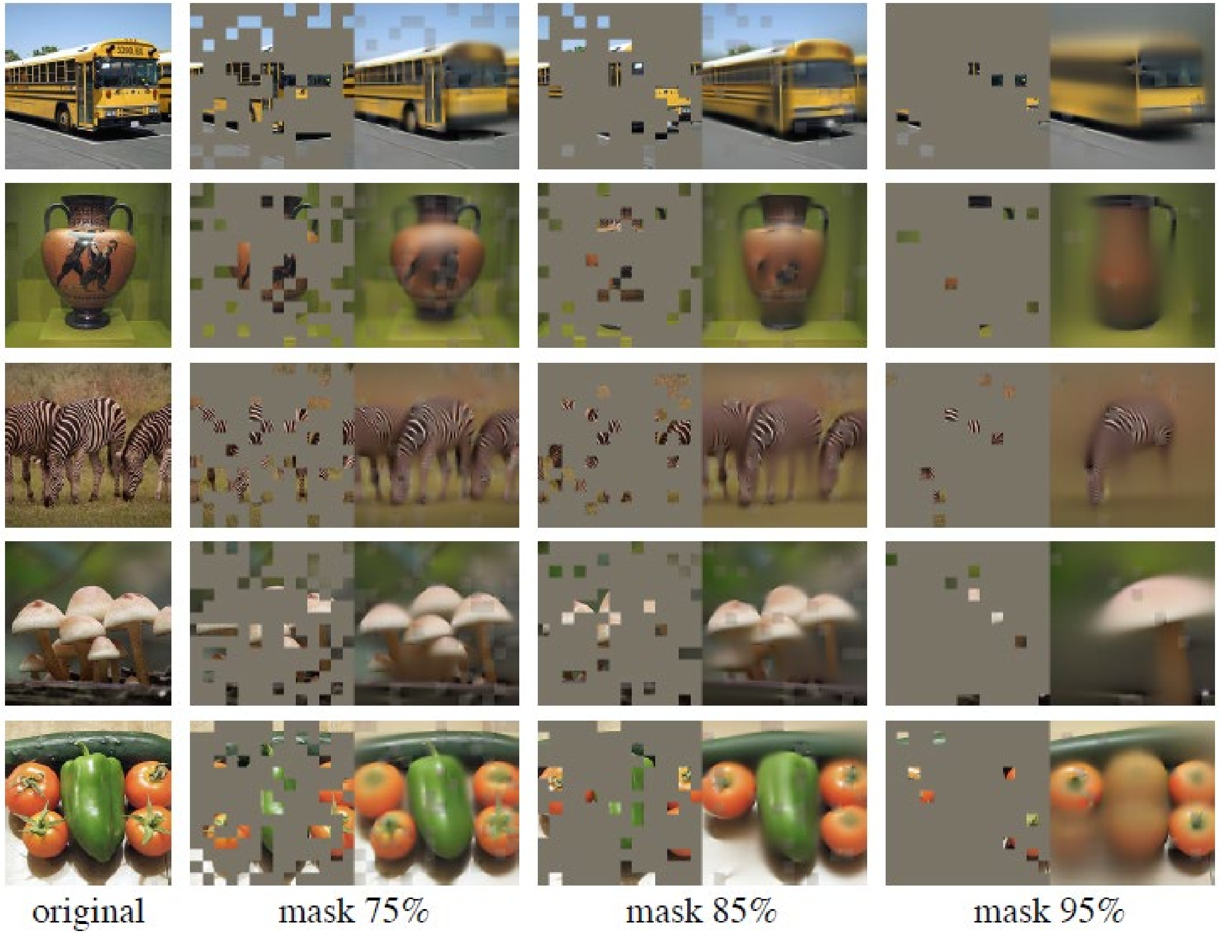


Figure 4. Reconstructions of ImageNet *validation* images using an MAE pre-trained with a masking ratio of 75% but applied on inputs with higher masking ratios. The predictions differ plausibly from the original images, showing that the method can generalize.

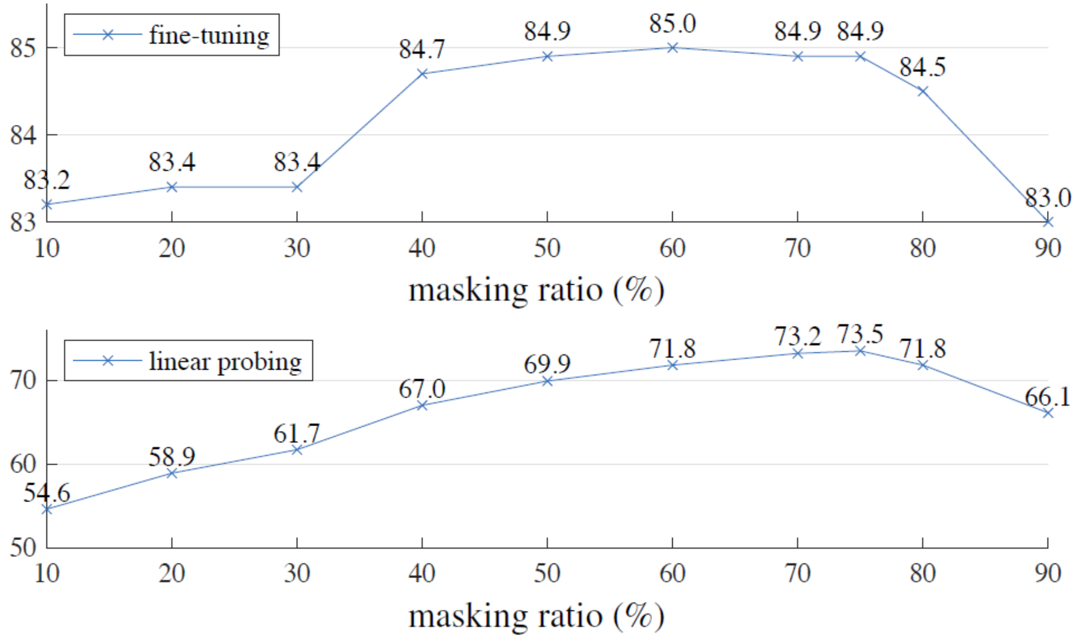


Figure 5. **Masking ratio**. A high masking ratio (75%) works well for both fine-tuning (top) and linear probing (bottom). The y-axes are ImageNet-1K validation accuracy (%) in all plots in this paper.

This is in contrast with BERT(15%) and also much higher than those in related works in vision (20%~50%)

All fine-tuning results are better that training from scratch (82.5%)

ImageNet Experiments

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).

The last layers in an autoencoder are more specialized for reconstruction

It has 8 bocks and a width of 512-d.
It only has 9% FLOPs per token vs. ViT-L (24 blocks, 1024-d)

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**×

Table 2. **Wall-clock time** of our MAE training (800 epochs), benchmarked in 128 TPU-v3 cores with TensorFlow. The speedup is relative to the entry whose encoder has mask tokens (gray). The decoder width is 512, and the mask ratio is 75%. †: This entry is estimated by training ten epochs.

ImageNet Experiments

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

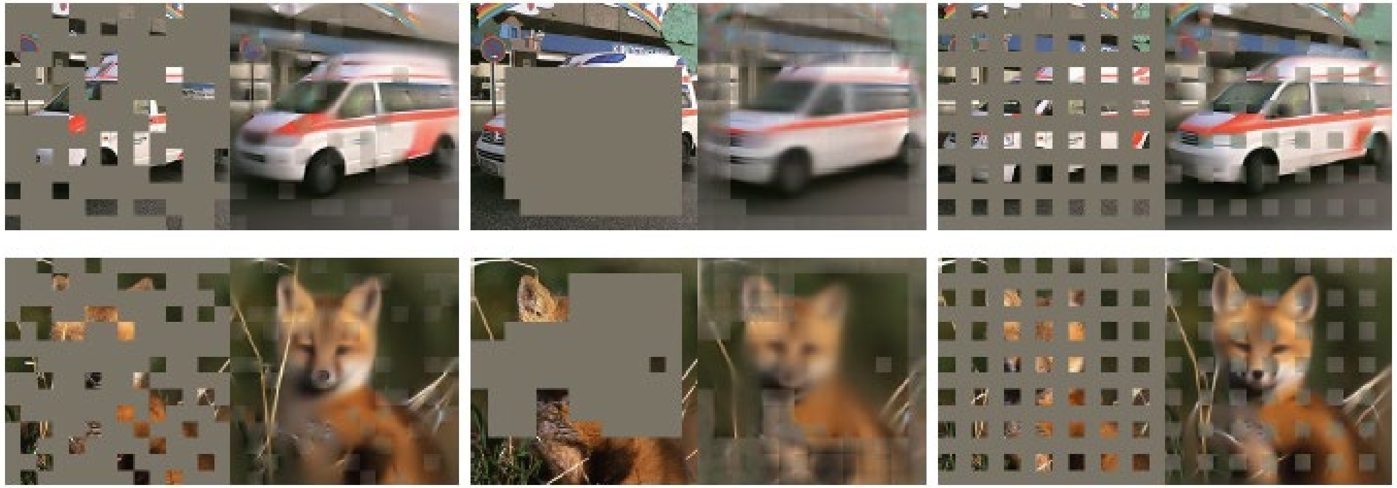
(e) **Data augmentation**. Our MAE works with minimal or no augmentation.

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.

The role of data augmentation is mainly performed by random masking

Training loss is higher
Reconstruction is also blurrier



random 75% block 50% grid 75%

Representation quality is lower

ImageNet Experiments

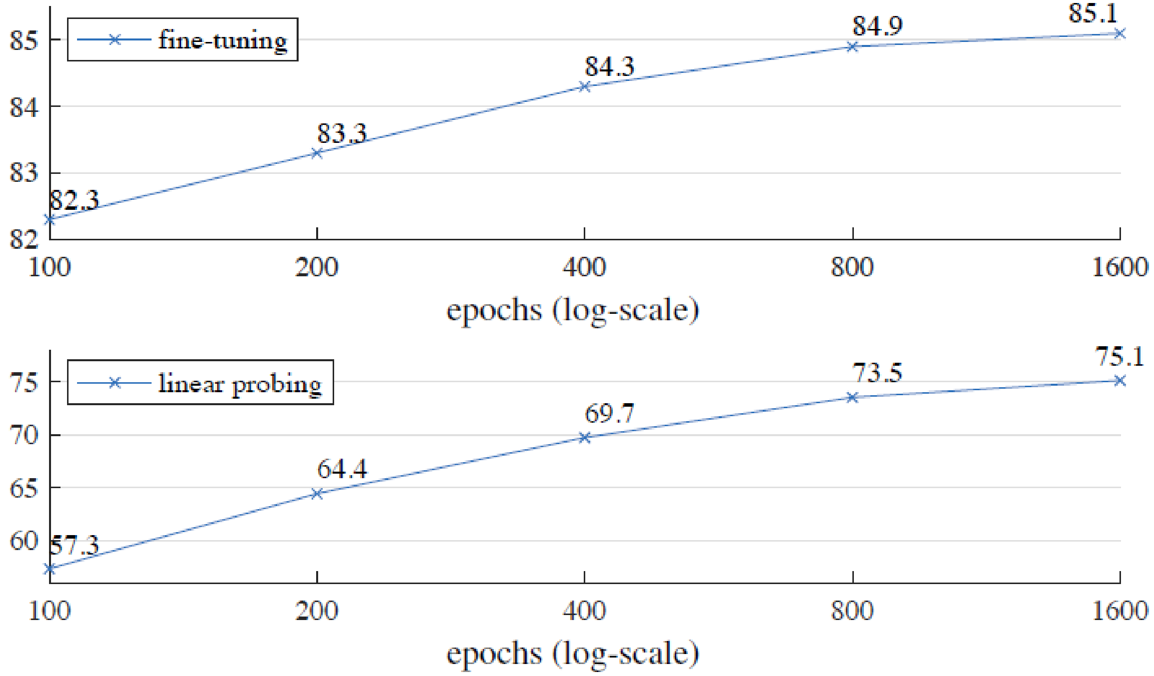


Figure 7. **Training schedules**. A longer training schedule gives a noticeable improvement. Here each point is a full training schedule. The model is ViT-L with the default setting in Table 1.

ImageNet Experiments

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.

Table 1. **MAE ablation experiments** with ViT-L/16 on ImageNet-1K. We report fine-tuning (ft) and linear probing (lin) accuracy (%). If not specified, the default is: the decoder has depth 8 and width 512, the reconstruction target is unnormalized pixels, the data augmentation is random resized cropping, the masking ratio is 75%, and the pre-training length is 800 epochs. Default settings are marked in gray.

Fine-tune과 Linear probing의 경향이 다름

It misses the opportunity of pursuing strong but non-linear features – which is indeed a strength of deep learning

We study a partial fine-tuning protocol: fine-tune the last several layers while freezing the others

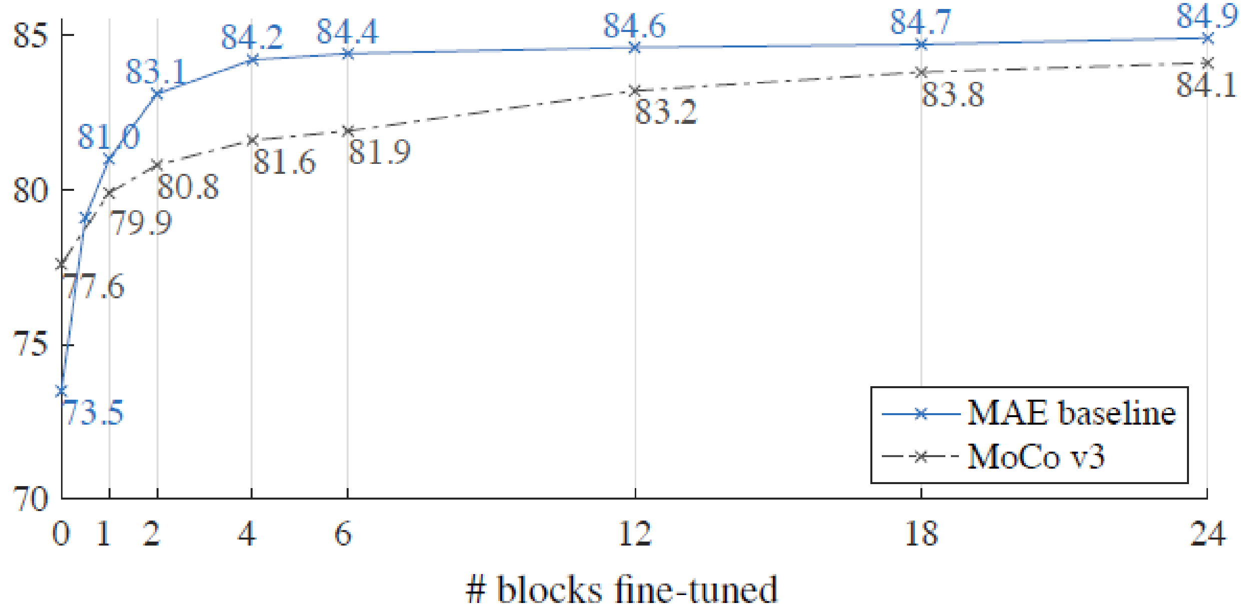


Figure 9. **Partial fine-tuning** results of ViT-L w.r.t. the number of fine-tuned Transformer blocks under the default settings from Table 1. Tuning 0 blocks is linear probing; 24 is full fine-tuning. Our MAE representations are less linearly separable, but are consistently better than MoCo v3 if one or more blocks are tuned.

MAE representations are less linearly separable, but they are stronger non-linear features and perform well when a non-linear head is tuned.
These observations suggest that linear separability is not the sole metric for evaluating representation quality

Transfer Learning Experiments

MI2L Medical Imaging Intelligent Reality Lab 울산대학교 UNIVERSITY OF ULSAN 서울아산병원 Asan Medical Center

method	pre-train data	AP^{box} ViT-B	AP^{box} ViT-L	AP^{mask} ViT-B	AP^{mask} ViT-L
supervised	IN1K w/ labels	47.9	49.3	42.9	43.9
MoCo v3	IN1K	47.9	49.3	42.7	44.0
BEiT	IN1K+DALLE	49.8	**53.3**	44.4	47.1
MAE	IN1K	**50.3**	**53.3**	**44.9**	**47.2**

Table 4. **COCO object detection and segmentation** using a ViT Mask R-CNN baseline. All entries are based on our implementation. Self-supervised entries use IN1K data *without* labels. Mask AP follows a similar trend as box AP.

method	pre-train data	ViT-B	ViT-L
supervised	IN1K w/ labels	47.4	49.9
MoCo v3	IN1K	47.3	49.1
BEiT	IN1K+DALLE	47.1	53.3
MAE	IN1K	**48.1**	**53.6**

Table 5. **ADE20K semantic segmentation** (mIoU) using UperNet. BEiT results are reproduced using the official code. Other entries are based on our implementation. Self-supervised entries use IN1K data *without* labels.

dataset	ViT-B	ViT-L	ViT-H	ViT-H_{448}	prev best
iNat 2017	70.5	75.7	79.3	**83.4**	75.4 [55]
iNat 2018	75.4	80.1	83.0	**86.8**	81.2 [54]
iNat 2019	80.5	83.4	85.7	**88.3**	84.1 [54]
Places205	63.9	65.8	65.9	**66.8**	66.0 [19]†
Places365	57.9	59.4	59.8	**60.3**	58.0 [40]‡

Table 6. **Transfer learning accuracy on classification datasets**, using MAE pre-trained on IN1K and then fine-tuned. We provide system-level comparisons with the previous best results.
†: pre-trained on 1 billion images. ‡: pre-trained on 3.5 billion images.

	IN1K			COCO		ADE20K	
	ViT-B	ViT-L	ViT-H	ViT-B	ViT-L	ViT-B	ViT-L
pixel (w/o norm)	83.3	85.1	86.2	49.5	52.8	48.0	51.8
pixel (w/ norm)	83.6	85.9	86.9	50.3	53.3	48.1	53.6
dVAE token	83.6	85.7	86.9	50.3	53.2	48.1	53.4
$\triangle$	0.0	-0.2	0.0	0.0	-0.1	0.0	-0.2

Table 7. **Pixels *vs.* tokens** as the MAE reconstruction target. $\triangle$ is the difference between using dVAE tokens and using normalized pixels. The difference is statistically insignificant.