
Masked Autoencoders are Scalable Vision Learners

School of Industrial and Management Engineering, Korea University

Eun Ji Koh

Contents

- ❖ Research Purpose
- ❖ Masked Autoencoders are Scalable Vision Learners
- ❖ Experiments
- ❖ Conclusion

Research Purpose

❖ Masked Autoencoders are Scalable Vision Learners (arXiv, 2021)

- Facebook AI Research (FAIR)에서 연구한 논문으로 2021년 11월 11일 공개

Masked Autoencoders Are Scalable Vision Learners

Kaiming He^{*,†} Xinlei Chen^{*} Saining Xie Yanghao Li Piotr Dollár Ross Girshick

^{*}equal technical contribution [†]project lead

Facebook AI Research (FAIR)

Abstract

This paper shows that masked autoencoders (MAE) are scalable self-supervised learners for computer vision. Our MAE approach is simple: we mask random patches of the input image and reconstruct the missing pixels. It is based on two core designs. First, we develop an asymmetric encoder-decoder architecture, with an encoder that operates only on the visible subset of patches (without mask tokens), along with a lightweight decoder that reconstructs the original image from the latent representation and mask tokens. Second, we find that masking a high proportion of the input image, e.g., 75%, yields a nontrivial and meaningful self-supervisory task. Coupling these two designs enables us to train large models efficiently and effectively: we accelerate training (by 3× or more) and improve accuracy. Our scalable approach allows for learning high-capacity models that generalize well: e.g., a vanilla ViT-Huge model achieves the best accuracy (87.8%) among

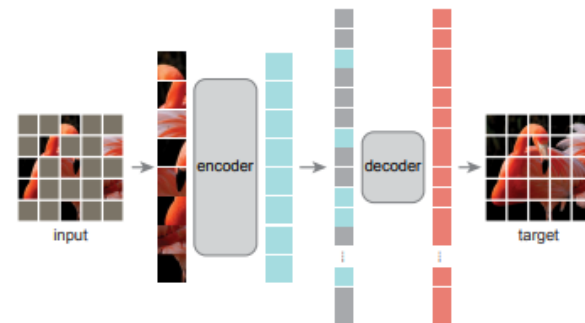


Figure 1. **Our MAE architecture.** During pre-training, a large random subset of image patches (e.g., 75%) is masked out. The encoder is applied to the small subset of *visible patches*. Mask tokens are introduced *after* the encoder, and the full set of encoded patches and mask tokens is processed 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 to produce representations for recognition tasks.

Research Purpose

❖ Motivation & Background

- 딥러닝이 발전함에 따라 model의 크기가 커지고 수 백만 개 이상의 data가 요구됨
- NLP 분야에서는 self-supervised pre-train을 활용하여 model의 크기를 성공적으로 확장
 - 대표적으로 GPT, BERT: data의 일부를 제거하고 이를 예측하고자 하는 모델
- Vision 분야에서는 masked autoencoding에 대한 progress가 NLP 분야에 비해 두드러지지 않음

Research Purpose

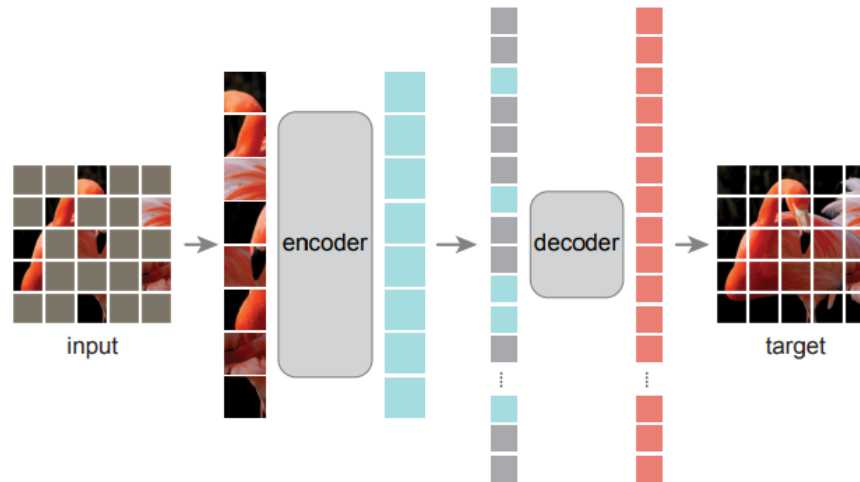
❖ Motivation & Background

- Vision 분야에서 masked autoencoding에 대한 progress가 뒤쳐지는 이유에 대한 저자들의 분석
 - 1) **Image와 text를 위한 model의 architecture 차이**
 - 기존 vision 분야에서는 regular grid에서 작동하는 convolution이 주로 사용되기 때문에 mask token이나 positional embedding을 network에 포함시키기 어려움
 - 최근 Vision Transformer(ViT)의 등장으로 해결
 - 2) **Image와 text가 갖는 information density 차이**
 - NLP 분야에서는 모델이 mask token 예측 task를 수행하며 언어에 대한 정교한 이해가 가능
 - Vision 분야는 heavy spatial redundancy로 인해 image의 high-level에 대한 이해 없이도 mask token을 복원 가능
 - 3) **Image와 text를 위한 autoencoder의 decoder 역할 차이**
 - NLP 분야에서는 decoder가 missing words를 예측하며 rich semantic information을 파악
 - Vision 분야에서는 decoder가 pixel을 복원하기 때문에 output이 low semantic level에 그침

Masked Autoencoders are Scalable Vision Learners

❖ Masked Autoencoders are Scalable Vision Learners: MAE

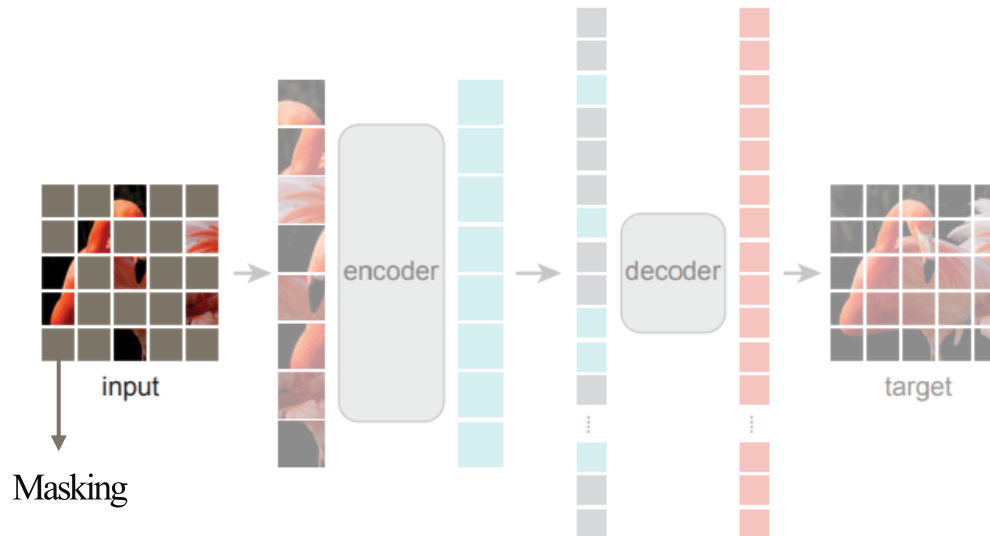
- 본 논문은 visual representation learning을 위해 scalable한 형태의 masked autoencoder를 제안
- 다음 방식으로 기존 vision 분야의 mask autoencoder가 갖는 문제를 해결
 - Mask token의 비율을 매우 높임으로써 모델이 image의 redundancy에 의존하지 않고 high-level semantic을 파악하도록 함
 - Encoder와 decoder를 asymmetric하게 design하여 decoder를 경량화 하였으며 쉽게 큰 모델로 확장 가능



Approach

❖ Masking

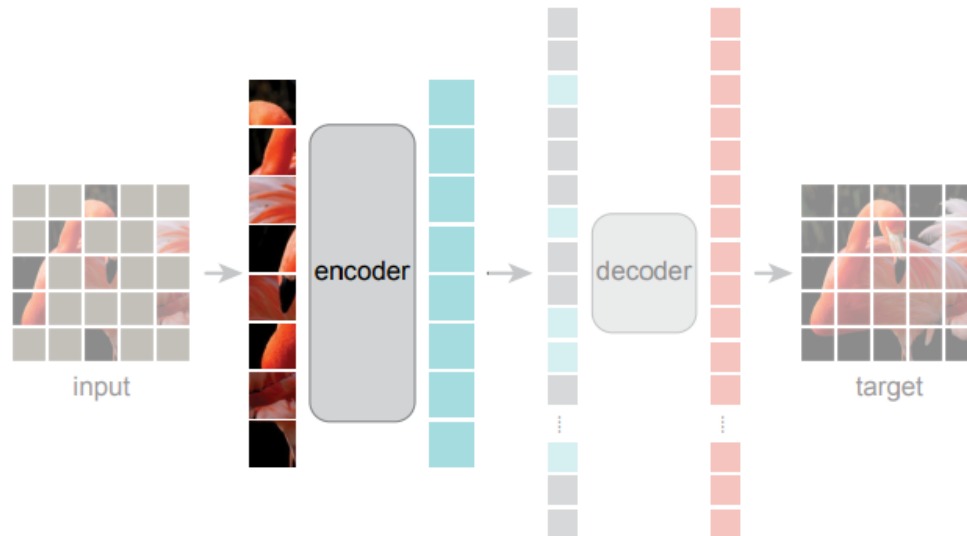
- Image를 regular non-overlapping patch로 분할(ViT와 동일)
- 이후 Uniform Dist.에 따라 random sampling(without replacement) 하여 masking
 - Image의 center에 위치한 patch 위주로 masking 되는 것을 방지하기 위함
- 전체 patch 중 높은 비율로 masking하여 redundancy를 제거
 - 인근 patch를 통한 image 복원이 아닌 image에 대한 high-level semantic 파악을 위함



Approach

❖ MAE encoder

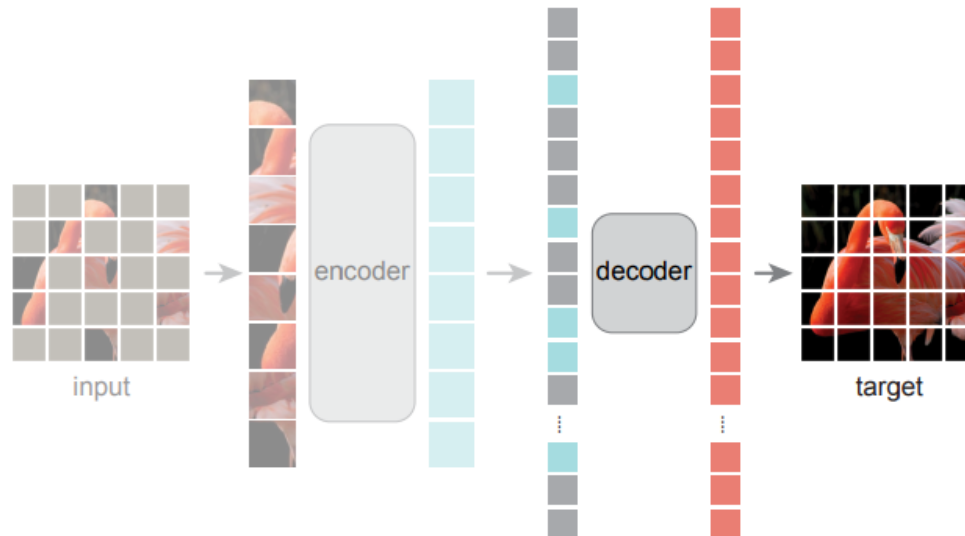
- Vision Transformer(ViT)와 동일 구조
- Mask token을 제외하고 linear projection을 통해 patch embedding
 - Masking 하지 않은 적은 수의 patch만을 활용하기 때문에 다량의 데이터도 효율적으로 학습 가능
- 이후 positional embedding 부여



Approach

❖ MAE decoder

- Encoder를 통과하여 생성된 latent representation과 mask token을 모두 사용 (asymmetric 구조)
 - Mask token: shared, learned vector로써 missing된 patch의 존재를 나타내는 역할만 수행
- Latent representation과 mask token에 positional embedding 부여
- Decoder는 원본 image로 복원하는 역할을 수행하며 pre-train 과정에만 사용
- Encoder에 비해 narrow, shallow한 구조 (lightweight)



Approach

❖ Reconstruction target & Simple implementation

- MAE는 mask token을 pixel 단위로 예측하여 input image를 복원하고자 함
- Loss function은 원본 image와 복원된 image 간의 MSE를 사용하며 mask token에 대해서만 계산 (BERT와 유사)
- MAE pre-train은 specialized sparse operation이 없어도 효율적으로 구현 가능

Experiments

❖ ImageNet Experiment

- 저자들은 ImageNet-1k를 사용하여 self-supervised pre-training을 진행
- 이후 representation 평가를 위해 1) end-to-end fine-tuning, 2) linear probing 사용
- Baseline: ViT-Large
 - ✓ 기존 ViT 논문에서 ViT-Large의 성능: scratch, original
 - ✓ 기존 ViT-Large의 성능을 끌어올리기 위해 저자들이 여러 노력한 후: scratch, our impl.
 - ✓ 본 논문의 MAE: baseline MAE

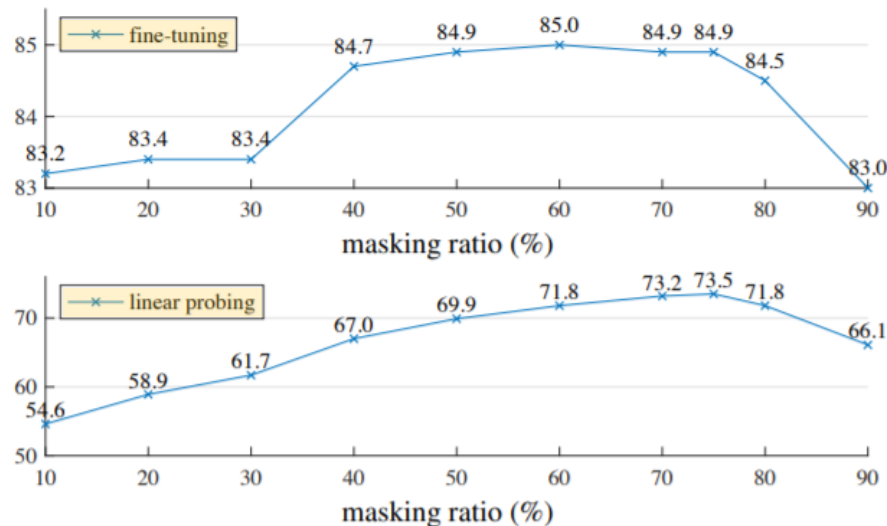
scratch, original [16]	scratch, our impl.	baseline MAE
76.5	82.5	84.9

ImageNet-1k에 대한 valid set의 acc score

Experiments

❖ Masking Ratio

- Optimal ratio가 매우 높음 (약 15% masking 을 하는 BERT와 다름)
 - Fine-tuning과 Linear probing 모두 75% masking 한 경우 성능이 가장 좋음
- Fine-tuning의 경우 masking ratio와 무관하게 항상 scratch로 학습한 모델(82.5)보다 성능이 좋음



Masking 비율에 따른 ImageNet-1K validation accuracy 변화

Experiments

❖ Decoder design

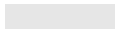
- MAE에서 사용한 decoder는 기존 ViT의 연산량의 9%에 해당하는 lightweight한 구조
- Decoder의 depth(block의 개수)의 측면
 - Linear probing의 경우 depth 증가에 따라 성능이 꾸준히 오르는 이유
: autoencoder의 마지막 일부 layer는 recognition 보다 reconstruction에 specialized되어 있기 때문
 - Fine-tuning의 경우 decoder의 depth에 따른 성능차이가 상대적으로 적음
: Fine-tuning을 통해 recognition task에 적합한 방식으로 학습이 가능해지기 때문

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

Decoder depth에 따른 Acc 비교

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

Decoder width에 따른 Acc 비교

- Linear probing : lin
- Fine-tuning : ft
-  : MAE의 기본 모델

Experiments

❖ Mask token

- MAE encoder에서 mask token을 사용하는 경우 성능이 저하됨
 - Encoder에서 mask token을 사용하여 학습하게 되면, inference 시에는 mask token이 없는 image를 사용하므로 train set과 test set의 괴리가 생기기 때문
- MAE encoder는 mask token을 사용하지 않기 때문에 타 모델에 비해 학습 속도가 빠르고 적은 메모리를 사용하기 때문에 큰 모델, large-batch training이 가능 (scalable)

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

Encoder에서의 mask token 사용에 따른 Acc 비교

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×

MAE의 wall-clock time

Experiments

❖ Reconstruction target & Data augmentation

- Pixel을 직접 예측하는 모델이 가장 좋은 성능을 보임
- Pixel의 평균, 표준편차를 활용하여 normalization 한 후 예측하도록 하는 것에서 가장 좋은 성능
- Making이 일종의 augmentation의 역할을 수행하는 것으로 간주할 수 있기 때문에 augmentation에 대한 의존도가 낮음

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

Data augmentation에 따른 Acc 비교

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

Reconstruction target에 따른 Acc 비교

- Pixel: pixel 예측
- PCA: PCA coefficient 예측
- dVAE token: token 예측

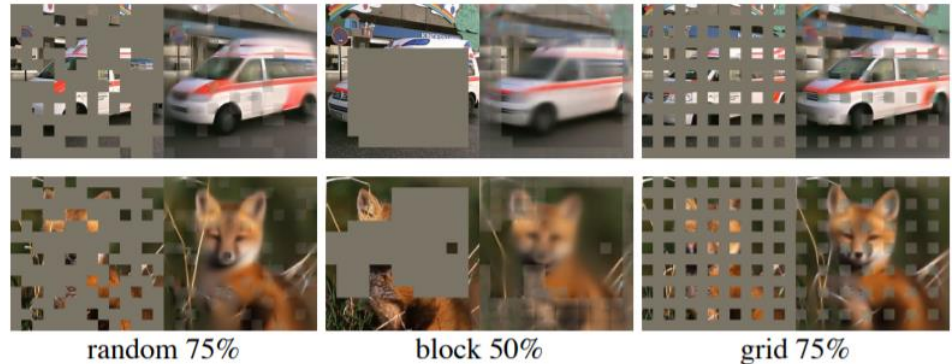
Experiments

❖ Mask sampling strategy

- Random하게 masking을 하는 경우에 가장 좋은 성능을 보임
- Block 방식의 경우 75%의 비율로 masking을 하면 성능이 저하됨
- Grid의 경우 image 복원이 잘 되는 것으로 보이나 representation의 성능이 좋지 않음 (linear probing)

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

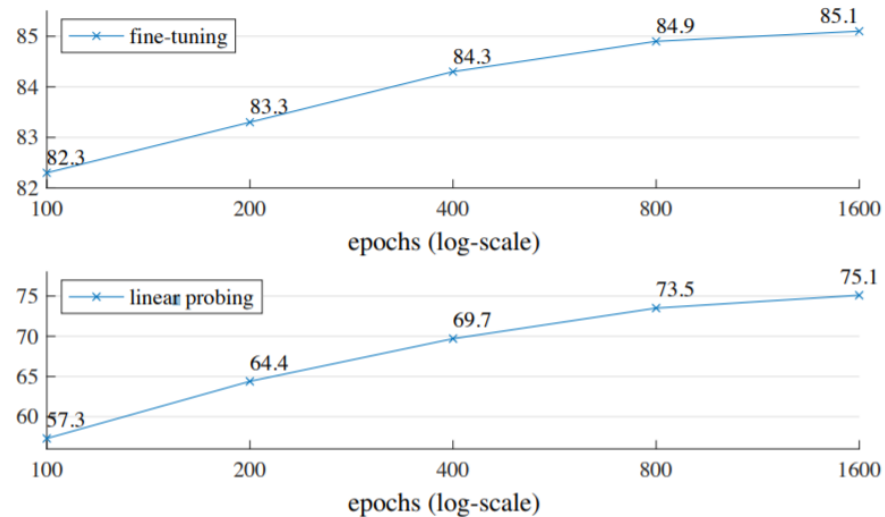
Data sampling 방식에 따른 Acc 비교



Experiments

❖ Training Schedule

- Training 시간이 증가함에 따라 성능도 향상됨
- Contrastive learning과 다른 양상
(ViT-L 모델을 사용하는 경우 MoCo v3는 300 epochs에서 saturate 됨)



Training 시간에 따른 Acc 비교

Experiments

❖ Comparisons with self-supervised methods

- 모델 규모가 커짐에 따라 성능 차이가 크게 벌어짐
- MAE는 scale up이 매우 잘되는 모델

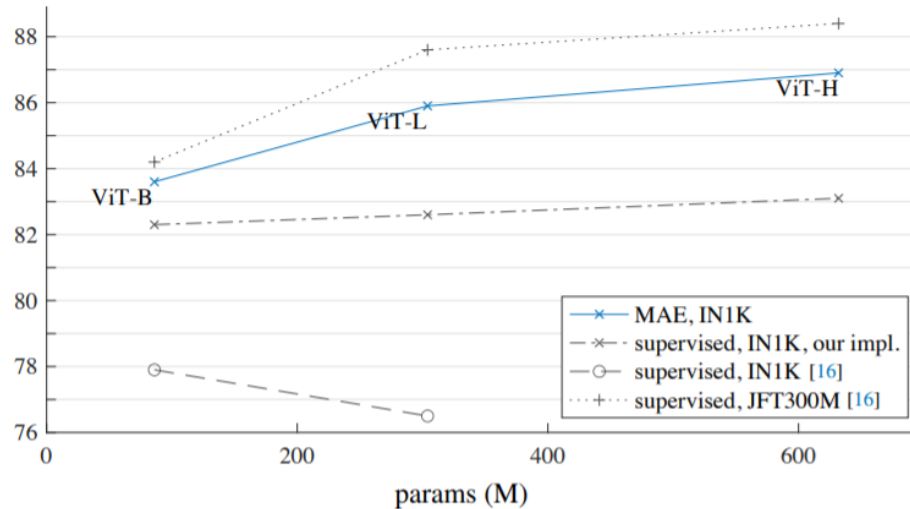
method	pre-train data	ViT-B	ViT-L	ViT-H	ViT-H ₄₄₈
scratch, our impl.	-	82.3	82.6	83.1	-
DINO [5]	IN1K	82.8	-	-	-
MoCo v3 [9]	IN1K	83.2	84.1	-	-
BEiT [2]	IN1K+DALLE	83.2	85.2	-	-
MAE	IN1K	<u>83.6</u>	<u>85.9</u>	<u>86.9</u>	87.8

타 self-supervised learning 방법론과의 Acc 비교

Experiments

❖ Comparisons with supervised pre-training

- 기존 ViT 모델의 경우 모델이 scale up 됨에 따라 성능이 하락하거나 더 이상 오르지 않음
- MAE 기반의 모델은 encoder의 크기만 scale up 됨에 따라 성능이 향상됨

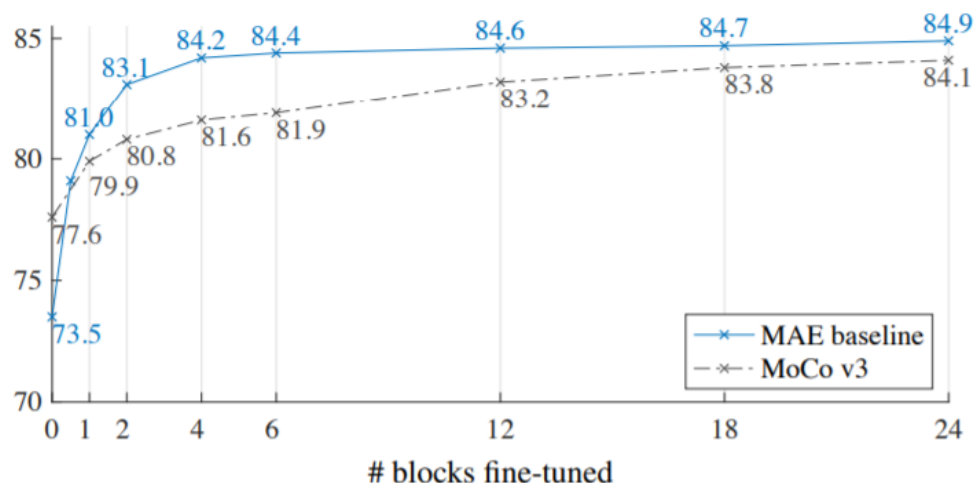


MAE pre-training과 supervised pre-training 간의 Acc 비교

Experiments

❖ Partial Fine-tuning

- 본 논문은 representation 평가를 위해 1) end-to-end fine-tuning, 2) linear probing 사용
 - Linear probing은 non-linear feature에 대한 opportunity를 miss하는 단점이 있음
 - 이를 해결하고자 last layer 일부만 fine-tuning 하는 partial fine-tuning protocol의 필요성 언급
- Linear probing보다 partial fine-tuning을 했을 때 성능이 향상됨



Fine-tuning block 개수에 따른 Acc 비교

Experiments

❖ Transfer Learning Experiments

- Object detection과 semantic segmentation에 대해 transfer learning 수행
- MAE가 다양한 down stream task에서 모두 좋은 성능을 보임

method	pre-train data	AP ^{box}		AP ^{mask}	
		ViT-B	ViT-L	ViT-B	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

COCO object detection and segmentation

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

ADE20K semantic segmentation (mIoU)

Conclusion

❖ Conclusion

- 본 연구는 ImageNet과 transfer learning에서 NLP와 유사한 간단한 self-supervised 방법인 autoencoder가 scalable한 benefit을 제공할 수 있음을 확인함
- MAE는 간단하고 효율적으로 complex, holistic reconstruction을 수행하며 image의 semantic을 학습함

Reference

1. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2021). Masked Autoencoders Are Scalable Vision Learners. arXiv preprint arXiv:2111.06377.
2. PR-355: Masked Autoencoders Are Scalable Vision Learners <https://youtu.be/mtUa3AAxPNQ>

Thank You