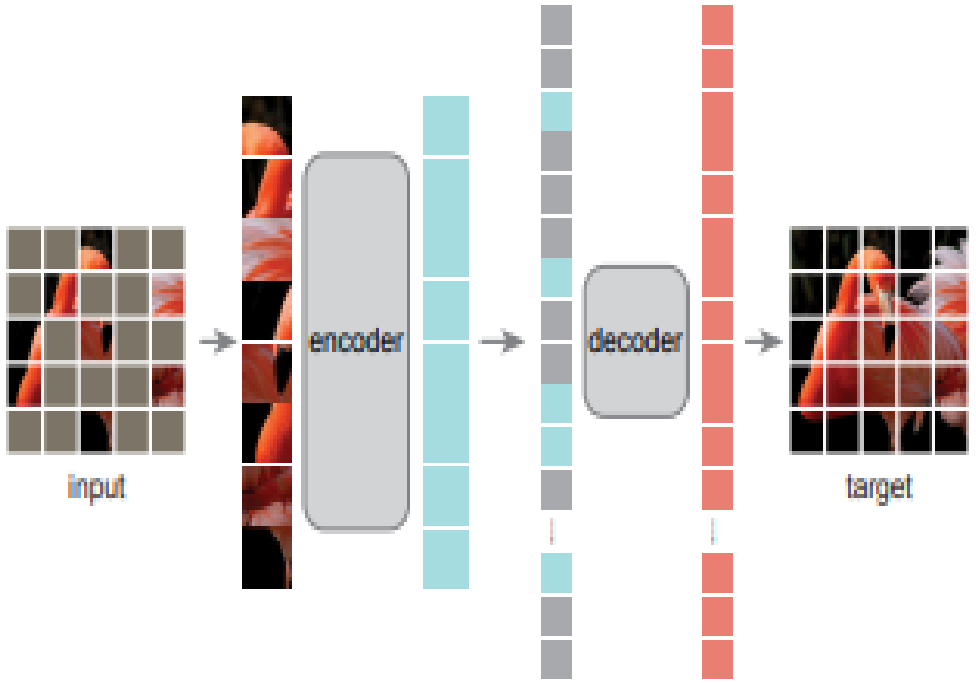
Masked Autoencoders are scalable Vision Learners

윤태영(인제대학교 헬스케어IT공학과 석사과정)

Introduction



MAE접근 방식(simple)

- 입력이미지를 패치로 나눠 random하게 masking을 하여 이를 재구성하고자 하는 모델
- 무작위 masking의 장점 : 중복 성이 적어지고, 전체적이해가 필요한 까다로운 self-supervised 학습에 적합.
- 재구성하는 경량 디코더와 마스킹 패치를 제외한 나머지에서만 작동하는 인코더 (비대칭 encoder-decoder)
- Input image의 75%를 마스킹할 시에 의미 있는 결과 (25%의 패치만 encoding하기때문에 계산 비용 3배 절약)

Related work

- **Masked language modeling(FERT, GPT)**
  - NLP의 pre-trained의 매우 성공적인 방법
  - input sequence 일부를 누락시키고 학습하여 누락 부분 예측
  - 사전 학습된 표현이 다양한 downstream task에 일반화.

- **Masked image encoder**
  - CNN을 사용하여 누락된 큰 영역을 inpainting함.
  - 최근에는 iGPT, BEiT와 같은 language model에 접목시켜 연구 진행.

- **AutoEncoder(classical method)**
  - DAE같은 경우 input을 noise로 손상시켜 손상되지 않은 raw signal처럼 재구성
  - MAE와 유사하면서 다른 method.

- **Self-supervised learning**
  - computer vision 또는 pre-training을 위한 다양한 사전 텍스트 작업에 초점.

Approach

- MAE는 부분 관찰을 통한 raw data 재구성 기법
- 부분 data를 encode를 통해 매핑하고, decode를 통한 raw data재구성
- 부분 data에 다른 비대칭 설계로 전체 재구성 → 일반적인 AE와 차별성을 둠.

- Masking
  - 겹치지 않게 이미지를 패치로 분할(like ViT)
  - 무작위 샘플링 : masking을 한 패치는 제거하고 나머지 패치들을 shuffle.

- MAE encoder
  - ViT와 같이 위치 임베딩을 가지지만 제거되지 않은 25%의 patch에만 적용

- MAE decoder
  - decoder는 pre-train시에만 사용
  - encoder에서 나온 결과와 masking되었던 패치와 합쳐 위치 임베딩을 추가함.
    (위치 임베딩이 없을 경우 자신의 위치에 대한 정보를 갖지 못하기 때문에 학습 불가)

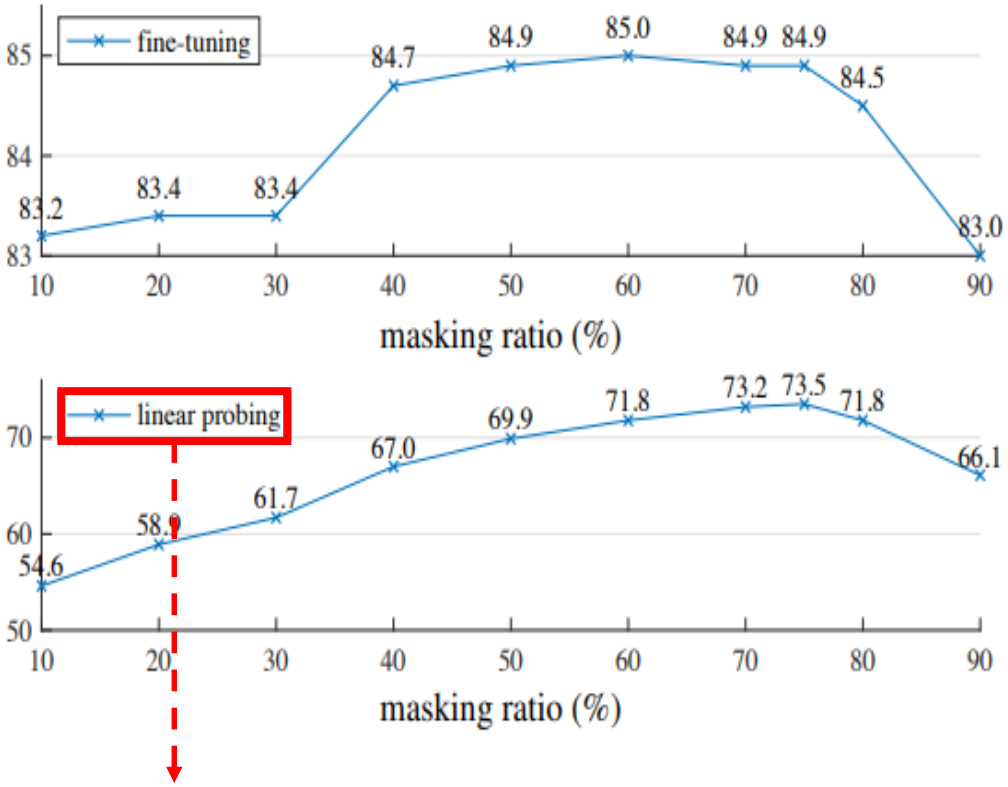
Approach

- **Reconstruction target**
  - Decoder 마지막 layer는 output channel 수와 patch들의 pixel수가 동일하다.(linear projection)
  - MSE : BERT와 유사하게 masked patch에만 손실 함수 계산

- **Simple implementation**
  - 특수한 sparse operation필요하지 않음.

    1. 모든 input patch에 토큰 생성
    2. 토큰을 임의로 섞고 비율에 맞춰 마지막 부분 제거(remove masked patch)
    3. encoding후에 masked 토큰과 encoder patch를 합침(None shuffle)
    4. decode는 전체에 적용(위치 임베딩 추가) →pre-train시에만 사용

Experiments - Main Properties



Linear probing은 다년간 유명한 protocol이지만,
비선형관계를 해결하지 못한다.

- Self-supervised learning을 통해 pre-train후에 fine-tuning과 linear probing을 할 수 있다.
- 두 method를 masking rate에 따른 변화를 보여준다.
- Masking rate은 40~80%정도에서 좋은 결과를 보임.
- Fine-tune과 linear probing의 추세가 다르고 fine-tuning이 더 좋은 결과를 보임.

Experiments - Main Properties

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).

A. Depth(Transformer 개수)의 변화

linear probing에서는 depth가 깊어질수록 성능개선이 보이지만 fine-tune에선 변화가 없음.

B. Width(Channel 개수)

512에서 가장 좋은 결과를 보이며, 비교 군인 ViT-L모델의 width가 1024인 점을 보았을 때 좀더 가볍다.

Experiments - Main Properties

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).

C. Encoder에서의 Mask token 유/무

w/o이 encoder에 mask token이 없는 것으로 mask token을 넣는 거보다 높은 결과를 보임.
또한, FLOPs를 보면 3.3배 차이가 남.

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.

block 50% grid 75%

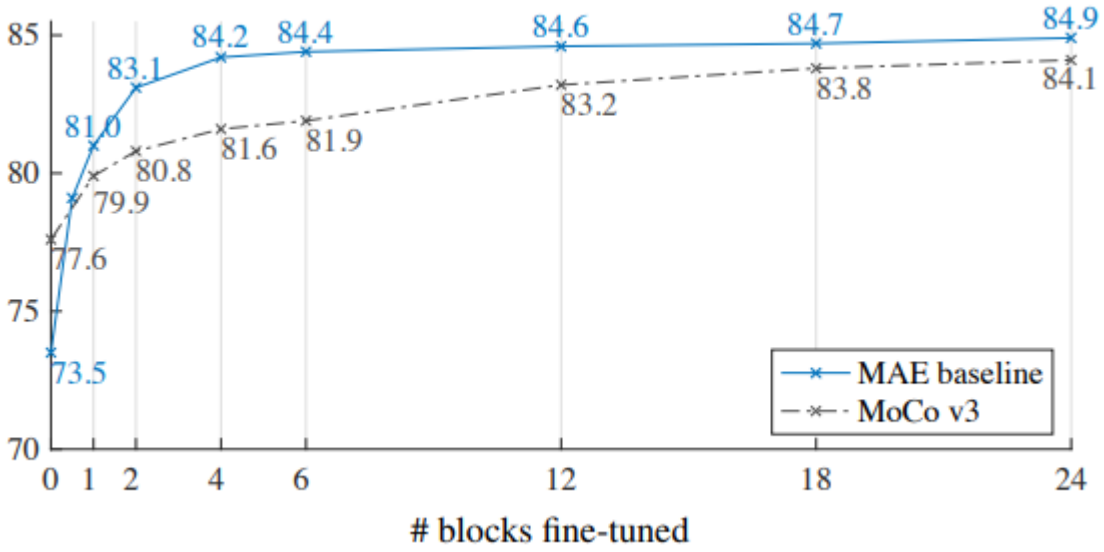
F. Mask sampling
block같은 경우 BEiT방식과 동일한 방법.
Block과 grid보다 random하게 마스킹을 했을 때가 더 좋은 성능을 보임.

Experiments – Comparison with 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**

- **Comparison with self-supervised methods**
  MAE는 ViT와 같이 Resolution을 통해 fine-tuning시 결과 향상에 도움이 됨.
  BEiT와 비교했을 때, MAE가 더 간단하고 빠르다.
  MAE training time : 31h (1600 epoch)
  Moco-V3 training time : 36h (300 epoch)

- **Partial Fine-tuning**
  partial fine-tune은 마지막 몇 개의 layer를 제외하고 freeze하는 방법이다.
  moco-V3는 linear probing에서는 좋은 결과를 보였지만, partial fine-tune에서는 결과가 좋지 않다.
  또한, MAE는 적은 layer의 fine-tune으로도 좋은 결과를 낸다.



Transfer learning experiments

		AP^{box}		AP^{mask}	
method	pre-train data	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

dataset	ViT-B	ViT-L	ViT-H	$ViT\text{-}H_{448}$	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]‡

- **Object detection and segmentation**
  MAE가 supervised learning에 비해 2.4~4.0포인트가 차이 날만큼 좋다.
  Pixel기반인 MAE와 token기반인 BEiT가 동등하거나 MAE가 더 좋지만 MAE가 더 빠른 성능을 가진다.

- **Classification datasets**
  모델이 커짐에 따라 성능이 증가.
  IN1K로 pre-train한 후 ViT-H에서 448로 resolution을 했을 때 성능향상이 된다.