Masked Autoencoders Are Scalable Vision Learners

https://arxiv.org/pdf/2111.06377

0. Introduction

- Vision Transformer(ViT) 기반 모델들은 대규모 데이터에서 강력한 성능을 보이나, Supervised Pretraining에 많은 라벨 데이터가 필요함.
- Self-Supervised Learning(SSL)은 라벨 없이 표현학습을 가능하게 하지만, 이미지 도메인에서 효율적인 SSL 구조가 제한적임.
- 본 논문은 Masked Autoencoder (MAE) 구조를 제안하여, 이미지에서 효율적이고 확장 가능한 Self-Supervised 학습을 목표로 함.
- 기여:
  1. ViT backbone과 자연스럽게 결합되는 Masked Autoencoder 구조 제안
  2. High masking ratio(예: 75%) 적용으로 효율적 학습
  3. Label 없이 사전 학습 후 다양한 Vision downstream task에서 성능 검증

1. Overview

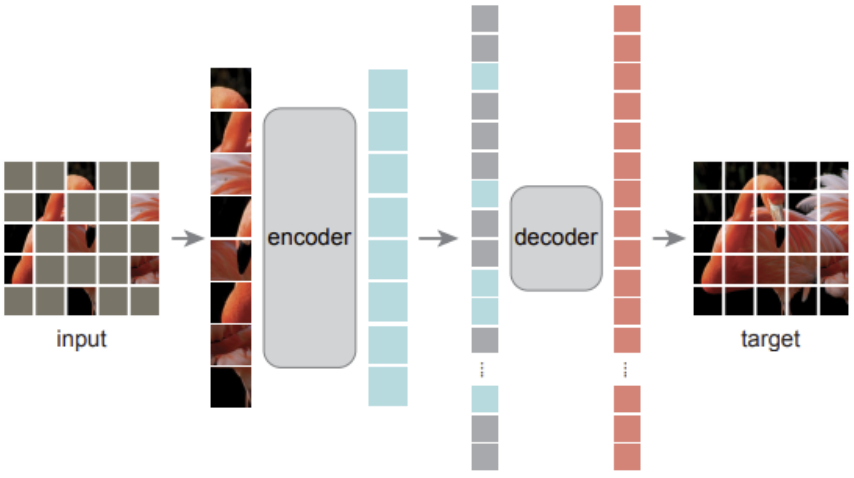
- 기본 아이디어: 입력 이미지를 Patch 단위로 분할하고, 주어진 patch의 일부(랜덤 75%)를 마스킹
- Encoder:
  - Visible patch만 입력으로 사용
  - ViT 기반 Transformer 구조
- Decoder:
  - Encoder 출력과 mask token을 모두 사용하여 원본 이미지 복원
- Pretraining → Fine-tuning 단계:

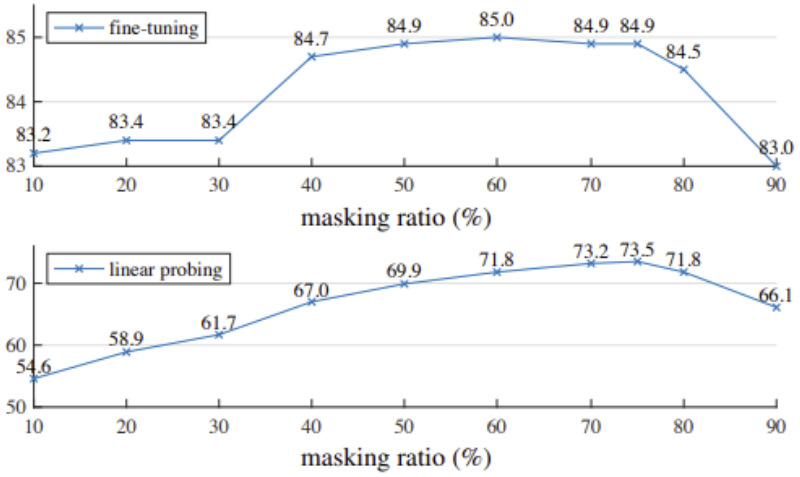
- Self-Supervised 단계에서 이미지 복원을 학습
- 이후 Fine-tuning을 통해 분류/탐지 성능 향상

2. Challenges

- 이미지 SSL에서 높은 성능을 얻기 위해서는:
  - 표현 다양성 확보
  - 계산 효율성 유지
  - Transformer 구조와 자연스러운 통합 필요
- 기존 Autoencoder 방식은 전체 입력을 처리해야 하므로 계산량 증가
- 높은 masking 비율에서 정보 손실 없이 좋은 표현 학습이 어려움

3. Method





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.

- Masking Strategy: 입력 이미지 patch 중 75%를 무작위로 masking, Encoder는 Visible patch만 계산하여 계산량 감소
- Encoder: Transformer 기반, 입력 patch embedding과 positional encoding
- Decoder: Lightweight Transformer, Mask token embedding 포함 전체 patch 복원
- Loss Function: Reconstruction loss(MSE), mask된 patch만 복원 오차 측정
- Training: Pretraining 단계에서 masking → encoder/decoder 학습, Fine-tuning 단계에서 classification 또는 downstream task 수행

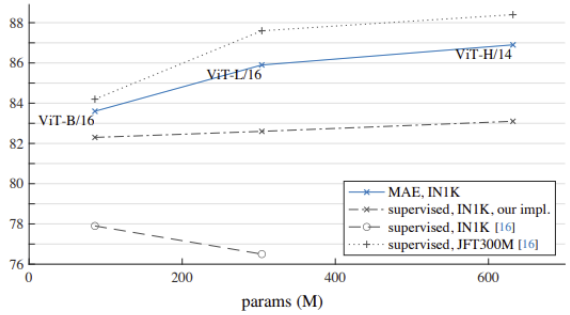
4. Experiments

- 데이터셋: ImageNet-1k (1.28M train images, 50k validation)
- Baseline 비교: ViT(Supervised), SimMIM, BEiT, iGPT

- 평가 지표: Top-1 Accuracy, Fine-tuning 성능
- 실험 설정: Encoder depth, masking ratio, decoder depth 실험, Fine-tuning 시 learning rate, batch size 조정

5. 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**



		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

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
△	0.0	-0.2	0.0	0.0	-0.1	0.0	-0.2

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

- MAE는 기존 Self-Supervised 방법 대비 성능 우위
- Masking ratio 75%에서 효율적 학습
- Decoder를 간소화하면서 계산량 절약
- Fine-tuning 성능 향상

6. Insight

- Key Strengths: Masking 기반 구조로 연산 효율 극대화, Label 없이 강력한 표현 학습 가능, 다양한 downstream task 확장 가능
- Limitations: High masking 비율의 정보 손실 가능성, Decoder 구조 설계 민감도 존재
- 후속 연구 방향: Multi-Modal MAE(이미지+텍스트), Video MAE(시계열 영상), Masking 전략 최적화 연구