MAE: Masked Autoencoders Are Scalable Vision Learners

기간	@05/14/2025 → 05/20/2025
주차	11주차
논문	https://arxiv.org/pdf/2111.06377
상태	완료
예습/복습	예습과제

0. Abstract

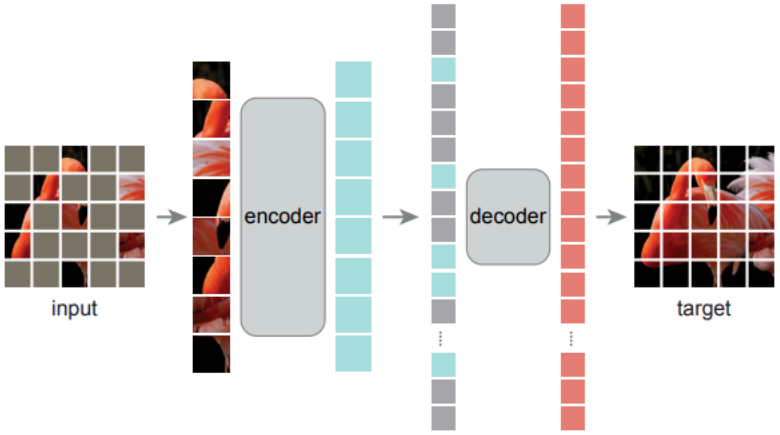
- **Masked autoencoders(MAE)** → 컴퓨터 비전을 위한 확장 가능한 self-supervised 학습 방법임을 보여줌.
- 접근 방식 simple!
    - 입력 이미지의 일부 패치를 무작위로 마스킹하고, 누락된 픽셀을 복원하도록 학습시킴.

    1. “**asymmetric 인코더-디코더 구조**”
        - 인코더는 마스킹되지 않은 패치(visible subset)만 처리.
        - 디코더는 잠재 표현(latent representation)과 마스크 토큰을 이용하여 원본 이미지 복원.
    2. “**입력 이미지의 상당 비율(ex. 75%)을 마스킹하는 것**”

    - nontrivial하고, 의미 있는 self-supervised 학습 과제 만듦.

  ⇒ 이 두 설계를 결합함으로써 학습 속도 3배 이상 향상 + 정확도 향상

- 고용량 모델 학습이 가능
  - vanilla ViT-Huge 모델 : ImageNet-1K만 사용한 경우 최고 정확도(87.8%) 달성
- 전이 학습(transfer learning) 에서도 supervised pretraining보다 나은 성능 보임, 확장성 측면에서도 유망한 패턴 나타냄.



1. Introduction

논문이 다루는 분야

- **Computer Vision**
- 특히, Self-Supervised Learning 방식으로 시각적 표현 학습(visual representation learning) 수행하는 방법론 개발&평가.

해당 task에서 기존 연구 한계점

1. 구조적 차이(Architectural Gap)
   - 과거) CNN 기반 네트워크가 주류, BERT 스타일의 마스크 토큰이나 위치 임베딩을 적용하기 어려웠음.
   - 최근 **Vision Transformer(ViT)** 도입으로 이 문제가 해결됨.
2. 정보 밀도 차이(Information Density Gap)
   - NLP → 고도로 의미 있는 신호(텍스트) 기반인 반면, 이미지 → spatial redundancy(공간적 중복성)이 많아 마스킹만으로 충분한 표현 학습 유도하기 어려움.

3. 디코더 역할의 차이(Decoder Functionality Gap)
    - NLP → 디코더가 의미 단위(단어) 예측
    - Vision → 저수준 픽셀 복원이므로 고수준 의미 표현 학습 어려움.

논문의 contributions

- **asymmetric 인코더-디코더 구조 도입**
    - 인코더는 visible 패치만 처리
    - 디코더는 마스크된 부분만 복원

    ⇒ 연산량 감소 및 효율성 극대화
- **고비율 마스킹 전략 제안**
    - 이미지 중 대부분을 마스킹하여 단순한 보간이 불가능한 self-supervised 학습 과제 생성
- **단순한 구조로도 고성능 달성**
    - 복잡한 토큰화 없이도 픽셀 기반 복원만으로 최고 성능 달성
- **전이 학습 및 확장성 검증**
    - 객체 검출, 분할 등 다양하나 downstream task에서 기존 supervised 사전학습보다 우수한 성능 입증.