< Masked Autoencoders Are Scalable Vision Learners >

0. Abstract

본 논문은 Masked Autoencoder (MAE)가 컴퓨터 비전 분야에서 확장 가능한 자가 지도 학습 기법임을 보인다. MAE는 입력 이미지에서 무작위로 패치를 masking하고, 해당 픽셀을 복원하는 작업 과정을 수행한다.

핵심 설계 요소는 총 2가지이다.

먼저, **비대칭 인코더-디코더 구조**를 가진다.

인코더는 마스크되지 않은 패치만 처리하고, 디코더는 가벼운 구조로, 인코더의 latent representation과 마스크 토큰을 사용해 이미지를 복원한다.

또, **고비율 마스킹 과정**을 통해 의미 있는 자기 지도 학습 태스크를 형성하며,

학습 속도를 3배 이상 가속하고 정확도도 향상시킨다.

이 방식은 대규모 모델을 효율적으로 학습하게 하며, 다양한 다운스트림 작업에서도 supervised pre-training보다 더 뛰어난 전이 성능을 보이며, 확장성 또한 뛰어나다는 게 특징이다.

1. Introduction

최근 딥러닝 모델의 성능은 아키텍처와 하드웨어의 발전에 힘입어 급속히 향상되었다. 하지만 대규모 모델은 수억 개의 레이블된 이미지를 필요로 하고 이로 인해 접근성이 떨어진다. NLP 분야에서 이러한 문제를 self-supervised learning으로 해결해 왔다.

이미지에는 공간적으로 중복된 부분이 많아, 일부 패치를 가리는 것만으로는 고차원의 의미 학습이 어렵다. 이를 해결하기 위해 매우 높은 비율의 마스킹으로 학습을 어렵게 만들어 더 풍부한 표현을 유도하였다.

이 과정에서 여러 설계적 차이가 유도되었다.

먼저, 아키텍처의 차이로는, 과거에는 CNN이 주류였지만, 최근에는 ViT의 도입으로 BERT와 유사한 아키텍처 적용이 가능해졌다.

정보 밀도에도 차이가 발생하였는데, 텍스트의 경우 정보가 밀도 있게 담겨 있으나, 이미지의 경우 정보가 밀도 있지 않아 더 많은 마스킹이 필요하다.

마지막으로 디코더의 역할의 차이가 존재한다. NLP의 경우 마스킹된 단어 예측하여 높은 수준의 의미를 학습하고, Vision의 경우 픽셀 복원하여 낮은 수준의 정보를 학습한다. 이와 같이 이미지 디코더의 설계가 학습 표현의 수준에 큰 영향을 끼쳤다.

논문은 새로운 방법으로서 **무작위 패치 마스킹과 비대칭 인코더-디코더 구조가 결합된 형태를 제안한다**. 인코더의 경우 마스크되지 않은 패치만 처리하여 연산량이 감소하고, 디코더의 경우 가볍게 구성되어 전체 이미지를 복원한다. 이 방식으로 효율성과 정확도를 모두 확보 가능하다.

새로운 방법을 적용함으로써 ViT-Large/Huge 모델을 ImageNet-1K에서 성공적으로 학습하였다.

또, 전이 학습에서도 기존의 supervised pre-training보다 우수한 성능을 가진다는 것을 확인할 수 있었다.