1. Introduction

Self-attention 기반 아키텍처 중에서 특히 Transformer는 자연어 처리(NLP) 분야에서 기본적인 모델이 되었다. 일반적으로 대규모의 텍스트 데이터셋으로 사전 학습을 한 뒤, 소규모의 특정 작업 데이터셋으로 미세 조정하는 방식을 사용한다.

Transformer는 계산 효율성과 확장성을 가지고 있어 1000억 개 이상의 파라미터를 가진 초대형 모델까지 훈련할 수 있게 되었고, 성능 향상 측면에서 아직 한계가 보이지 않다.

반면에, 컴퓨터 비전(CV) 분야에서는 여전히 합성곱 신경망(CNN) 기반 아키텍처가 주로 활용이 된다. 최근 NLP의 성공을 보고 컴퓨터 비전 분야에서도 self-attention을 CNN과 결합하거나 아예 대체하려는 시도들이 등장하였다

그러나 이러한 self-attention 기반 모델들은 특수한 attention 패턴을 사용하기 때문에 현대의 GPU, CPU와 같은 하드웨어 가속기에서는 효과적으로 활용되지 못했다. 그래서 대규모 이미지 인식에서는 아직까지 여전히 전통적인 ResNet 기반의 아키텍처가 최고 성능을 보인다.

본 논문은 Transformer를 거의 그대로 이미지에 직접 적용하는 실험을 진행하였다.
구체적으로, 이미지를 작은 patch들로 나누고, 각 patch를 선형 임베딩한 이후 Transformer에 입력한다. 즉, 패치를 NLP의 단어 토큰처럼 이미지를 패치 시퀀스로 다루는 것이다.
본 논문은 이를 감독 학습 방식으로 이미지 분류 과정을 훈련하였다.

ImageNet과 같은 중간 크기의 데이터셋에서 강한 정규화 과정 없이 훈련하면, 이 모델들은 비슷한 크기의 ResNet보다 약간 낮은 정확도를 보였다.

그 이유는, Transformer가 CNN에 비해 전이 불변성이나 국소성과 같은 귀납적 편향이 부족하기 때문이다. 그래서 데이터가 적으면 일반화 성능이 떨어지는 것이다.

- ➔ Transformer는 이미지 구조를 잘 반영하지 못하여 데이터 수가 많지 않으면 성능이 약하다
- ➔ 이미지 구조란? 물체가 약간 움직여도 동일한 물체라고 인식하는 성질

그러나 1,400만~3억 장 이미지를 가진 더 큰 데이터셋으로 훈련하면 상황이 다르다.
이 과정에서 대규모의 훈련이 귀납적 편향보다 중요하다는 결과를 얻을 수 있었다.

본 논문의 모델인 **Vision Transformer**는 ImageNet-21k, JFT-300M과 같은 **대규모 데이터셋으로 사전 학습한 경우**, 소규모의 데이터셋 전이 학습에서도 뛰어난 성능을 보였다.

- ImageNet: 88.55%
- ImageNet-ReaL: 90.72%
- CIFAR-100: 94.55%
- VTAB 19개 작업 평균: 77.63%