💚

ViT: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

기간	@04/23/2025 → 04/29/2025
주차	8주차
논문	https://arxiv.org/pdf/2010.11929
상태	진행 중
예습/복습	예습과제

0. Abstract

- 트랜스포머(Transformer) 아키텍처 → 자연어 처리(NLP) 처리 분야에서 표준으로 자리 잡았지만, 컴퓨터 비전에서는 그 적용이 제한적.
    - 비전 분야에서는 **어텐션(Attention)** 기법이 CNN과 함께 사용되거나, CNN의 일부 구성요소만 대체하는 방식으로 사용됨.
- 본 논문에서는 CNN에 대한 의존이 **필수가 아님**을 보여줌.
    - 이미지 패치를 시퀀스로 간주, 이를 **순수 트랜스포머에 직접 적용**하는 방식으로 이미지 분류 작업에서 우수한 성능 낼 수 있음을 증명.
- 대규모 데이터셋에서 사전학습(pre-training) 후, 중소규모의 이미지 인식 벤치마크들(ex. ImageNet, CIFAR-100, VTAB 등)로 전이학습.

- **Vision Transformer(ViT)**는 기존의 최첨단 CNN 모델들과 비교하여 **우수한 성능** 보이면서 학습에 필요한 **계산 자원 훨씬 적음**.

1. Introduction

논문이 다루는 분야

- **컴퓨터 비전(Computer Vision)**, 특히 이미지 분류(Image Classification) 분야에서의 트랜스포머(Transformer) 모델의 활용 가능성 탐구.
- 기존 NLP에서의 트랜스포머 성공 사례를 이미지 처리로 확장하려는 시도 → **Vision Transformer(ViT)** 모델 제안, CNN에 대한 대안적 접근 실험함.

해당 task에서 기존 연구 한계점

- **CNN 중심 구조의 한계**
  - 기존 컴퓨터 비전 연구는 주로 CNN 기반으로 이루어졌으나, 이들은 지역성(locality), 이웃 구조, translation equivariance 등 강한 **inductive bias** 내재함.
    - 지역성 → 전역 정보(global context) 빠르게 파악하기 어려움. 초기 층에서는 먼 영역 간의 상호작용 없음.
    - 이웃 구조 → 이웃한 픽셀 간의 연관성. 서로 가까운 픽셀들이 더 관련 있다는 공간적 가정. 그러나 모든 문제에서 이 가정이 유효하다고 할 수 없음.
    - Translation Equivariance → 어느 위치에서 보더라도 같은 특징 인식할 수 있음.(필터 공유되기 때문) 그러나, 위치 정보 자체는 보존되지 않기 때문에 위치 기반 판단에는 불리함.
  - 이는 작은 데이터셋에서의 일반화에는 유리하지만, 확장성과 범용성에서는 한계를 보임.
- **트랜스포머 적용의 어려움**
  - 이전 연구에서는 트랜스포머를 이미지에 직접 적용하는 대신, CNN과의 결합 또는 일부 구성 요소의 대체에서 그쳤음.
  - 완전한 트랜스포머 기반 구조는 이론적 가능성은 있었지만, 하드웨어 최적화와 확장성 측면에서 효과적인 구현 어려웠음.
- **데이터 크기 제약**
  - 트랜스포머는 선천적으로 inductive bias가 적어, 소규모 데이터셋에서는 성능 저하가 두드러짐.

논문의 contributions

- **순수 트랜스포머 구조를 이미지에 직접 적용**
  - CNN 없이, 이미지를 고정 크기 패치로 나누고 이를 1D 토큰 시퀀스로 변환하여 트랜스포머에 입력.
  - BERT와 유사하게 `[class]` 토큰 사용하여 분류 작업 수행함.
- **대규모 사전 학습으로 성능 극대화**
  - ImageNet-21k 및 JFT-300M과 같은 대규모 데이터셋에 대한 사전학습 통해 ViT 성능 크게 향상.
  - 결과적으로 CNN 기반 SOTA 모델들을 효율성(적은 연산량)과 성능 측면에서 초월함.
- **모델 확장성 및 효율성 분석**
  - 다양한 ViT 모델 규모(ViT-B/L/H) 및 패치 크기를 실험.
  - 전통적인 CNN 대비 더 나은 성능-연산 효율 trade-off를 입증함.
- **전이 학습 및 소량 데이터 학습에도 우수한 성능**
  - 다양한 벤티마크(ImageNet, CIFAR-100, VTAB 등)에서 전이 학습 성능 우수.
  - 소량 데이터 상황에서도 기존 모델 대비 경쟁력 있는 성능 확인.