Vision Advanced Assignment

Vision Transformer

:이미지 분류에 transformer 구조를 성공적으로 적용시킨 논문. 2021년 ICLR 발표 이후, 현재 ViT 관련 많은 후속 연구 진행 중.

: ViT는 inductive bias가 약한 대신, 상대적으로 통합적이고 일반적인 모델 구조임. inductive bias가 약하여 일반화 성능이 떨어지는 문제를 data-driven training, 즉 매우 큰 데이터셋에서의 사전학습을 통해 극복.

1. **Introduction**

: Transformer 구조의 NLP task에서의 성공

1) 높은 computational efficiency(연산 효율성), scalability(확장성)

→ 데이터셋과 모델 크기가 계속 커져도 모델 성능이 포화되지 않고 지속적으로 개선.

→ 지속적으로 모델을 키울 수 있다는 건 높은 연산 효율성 없이는 이루어질 수 없음. 또한, 데이터셋만 추가적으로 주어지면 지속적으로 성능 향상이 이루어진다는 점에서 transformer 구조가 높은 확장성을 갖고 있음을 알 수 있음.

→ 그러나 computer vision task에서는 제대로 적용이 이뤄지지 않음.

2) Self-Attention을 추가한 CNN계열 모델 구조와 아예 convolution을 attention으로 대체한 구조 제안. 이론적으로 효율성은 보장되었지만 실제 하드웨어 가속기와의 호환성 떨어짐.

3) **Vision Transformer(ViT)**

- 이미지 자체를 **standard Transformer**에 직접적으로 넣어주는 형태로 모델 구성
- 이미지를 패치 단위로 쪼개 토큰화.
- 패치에 linear embedding 적용 후, 시퀀스로 만들어준 형태로 Transformer 입력으로 들어가게 됨.
- 중간 사이즈의 데이터셋에 학습을 시킬 경우, 별도의 강력한 regularization 없이 기존 ResNets 대비 훌륭한 성능을 보여주지는 못함.
- 이는 Transformer 구조 자체가 CNN 구조에 비해 inductive bias가 부족하여 많은 양의 데이터 없이는 일반화가 제대로 이루어지지 않았을 것으로 추측.
- 그러나 큰 사이즈의 데이터셋(1400만 ~ 3억장)에서 학습을 시킬 경우, 위의 구조적 한계 (lack of Inductive Bias)를 극복 가능함.

- ViT는 충분한 크기의 데이터셋에서 사전학습 후, 보다 적은 데이터셋을 가진 task에 전이 학습을 시킬 때 좋은 성능을 보여주었음.

2. CNN과의 차이점을 중심으로 REVIEW

1) Architecture 구조

- CNN: 주로 합성곱 및 풀링 레이어로 구성, 이미지의 지역적 패턴을 추출하기 위해 커널 사용.
- ViT: 비전 트랜스포머 아키텍처를 사용하여 이미지 처리. 이는 시퀀스 데이터를 다루는 데 사용되던 트랜스포머 모델을 이미지에 적용한 것.

2) 입력 데이터

- CNN: 일반적으로 고정된 크기의 이미지를 입력 받음.
- ViT: 이미지를 패치(patch)로 나누어 처리하며, 이 패치를 임베딩하여 트랜스포머에 입력으로 사용함. 즉, 입력 이미지 크기에 민감하지 않음.

3) 특성 추출

- CNN: 계층적으로 이미지의 로컬한 특성을 추출하고 이를 다양한 합성곱 레이어를 통해 조합하여 전역적인 특성을 얻음.
- ViT: 트랜스포머를 사용하여 패치 간의 관계를 모델링하고, 이를 통해 이미지의 전역적인 정보를 추출함.

4) 학습 방법

- CNN: 주로 역전파(backpropagation)를 사용하여 학습되며, 데이터 양과 모델 크기에 따라 많은 매개변수가 필요할 수 있음.
- ViT: 비지도 사전학습(Unsupervised Pretraining)과 지도 학습(Supervised Fine-tuning)을 결합한 전이학습(Transfer Learning) 방식을 사용하여 소량의 라벨된 데이터로도 뛰어난 성능을 달성할 수 있음.

5) 활용분야

- CNN: 이미지 분류, 객체 검출, 분할 등의 고전적인 컴퓨터 비전 작업에서 널리 사용됨.
- ViT: 이미지 분류를 비롯한 다양한 컴퓨터 비전 작업에서 사용되며, 텍스트와 함께 모두 트랜스포머를 활용하는 다중 모달(Multimodal) 작업에도 적용됨.

3. 요약

2021년 발간된 ViT 논문은 CNN과 다르게 트랜스포머 아키텍처를 사용하여 이미지를 처리하며, 패치 기반의 입력과 비지도 사전학습을 통한 특성 추출 방식 등이 주요 차이점이다. 이러한 차이로 인해 ViT는 적은 데이터로도 높은 성능을 달성할 수 있으며, 다양한 컴퓨터 비전 작업에서 활용되고 있다.