AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

* Transformers와 같은 self-attention based 아키텍처는 NLP에서 자주 사용되는 모델이 되었음. 주요한 접근은 거대한 text corpus를 pre-train하고, 그 다음 더 작은 task-specific한 데이터셋에 fine-tune하는 것.

* 그렇지만 컴퓨터 비전에서는 convolutional 아키텍처가 여전히 지배적임. NLP에서의 성과로 인해 self-attention과 CNN-like 아키텍처를 합치거나 convolution을 완전히는 대체하려는 시도가 있었음. 그러나 후자의 경우, specialized attention 패턴의 이용으로 현대 하드웨어 가속기에서 잘 scale 되지 않았음. 따라서 큰 범위의 이미지 인식에서 classic ResNet-like 아키텍처가 여전히 SOTA임.

* 논문에서는 NLP에서의 성공에 영향을 받아, 표준 트랜스포머를 가능한 적은 수정으로 직접적으로 이미지에 적용하는 것을 실험함

* 관련 연구

Transformers

- 이미지에 self-attention을 한 naïve한 적용은 각 픽셀이 모든 다른 픽셀에 attend 하는 것을 요구

- 엄청난 많은 수의 픽셀로, 이것은 현실적인 입력 사이즈의 범위에 맞지 않음

Cordonnier

- 입력 이미지에서 2 X 2 사이즈의 패치들을 추출함 -> 오직 작은 해상도의 이미지에게만 적용 가능함

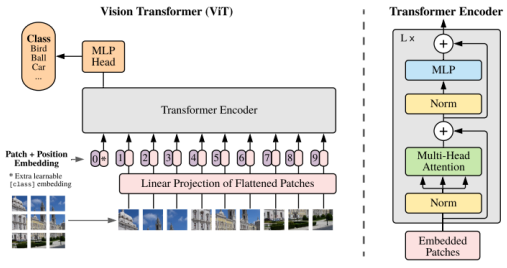
- 이 모델은 ViT와 유사함. 그렇지만 ViT는 거대한 스케일의 pre-training이 vanilla transformer가 SOTA CNN들과 경쟁력이 있다는 것을 보여줌 -> 중간 해상도의 이미지를 잘 처리함

Image GPT

- 이미지 해상도와 색 space를 줄인 뒤 트랜스포머를 이미지 픽셀들에 적용함

- generative 모델로 비지도 학습으로 훈련됨. 그리고 결과는 분류 문제를 위해 그 다음에 fine-tune되거나 linearly 조사함

* Model



- 이미지를 fixed-size 패치로 나눔 -> 각각을 linear하게 embed 함 -> 위치 임베딩을 더함 -> 표준 트

랜스포머 인코더에 결과로 나오는 시퀀스 벡터를 더함

- 분류를 수행하기 위해, 추가 학습이 가능한 토큰 분류를 시퀀스에 더하는 표준적인 접근을 이용함

- Vit(Visual Transformer)를 큰 데이터셋을 이용해 pre-train 한 다음, 작은 task에 맞게 fine-tune함

* Inspecting Vision Transformer

- Vision Transformer가 어떻게 이미지를 처리하는지 이해하기 위해 그것의 internal representation을 분석

: ViT의 첫번째 레이어는 flattened patch들을 더 낮은 차원의 공간으로 linear하게 투영함

-> 투영 이후에, 학습된 위치 임베딩은 패치 representation에 더해짐

- Self-attention은 ViT가 가장 낮은 레이어에 있는 전체 이미지에서의 정보를 통합하는 것이 가능하게 함

* Conclusion

- 컴퓨터 비전에서 self-attention을 사용한 이전 실험들과 달리, 초기 패치 추출 단계를 제외하고 image-specific inductive bias들을 아키텍처에 소개하지 않음

-> 이 간단한, 그렇지만 확장/축소가 가능한 전략은 pre-training에서 거대한 데이터셋과 연결할 때 놀랍게도 잘 됨

- ViT는 상대적으로 pre-train하기 저렴하면서, 이미지 분류 데이터셋에서의 SOTA와 아주 비슷하거나 초과함

- 여전히 해야 될 일들이 있음

1) ViT를 다른 컴퓨터 비전 task에 적용하는 것(예를 들면 detection, segmentation 등)

2) self-supervised pre-training를 계속 탐구하는 것

3) Further scaling of ViT