3주차: An Image is Worth 16x16 Words: Transformers for image recognition at scale

link: https://arxiv.org/pdf/2010.11929

Vision Transformer (ViT)

Abstract

트랜스포머는 기존에 자연어 처리의 기준이 되었지만 CV 쪽에는 잘 적용되지 않음

vision 분야에서는 convolutional network처럼 사용하는 정도로만 사용됨

대규모의 작은 이미지에 ViT를 학습시키면 더 적은 컴퓨터 자원을 사용하면서 좋은 결과를 보임

1. Introduction

Self-attention 기반의 아키텍처는 자연어 처리 분야에서 주로 쓰이는 모델로 large text corpus 로 사전 학습 후 작은 태스크 중심 데이터셋에 전이학습 시키는 접근법 사용

트랜스포머 기반 모델은 연산이 효율적이고 입력 시퀀스 길이에 영향받지 않아 확장성이 좋다

트랜스포머의 scaling success를 이미지에 적용할 수 있는 방법 연구.

이미지를 이미지 패치로 쪼개고, 최대한 수정을 덜해서 linear embedding으로 시퀀스 생성하고 트랜스포머의 input으로 활용해 이미지 분류를 지도학습함

중간 사이즈 데이터셋에 ViT 학습시키면 ResNet 모델에 비해 정확도가 약간 낮은 결과를 보임.

트랜스포머는 inductive bias가 없어 데이터 양이 불충분할 때 잘 일반화되지 않음

2. Related Work

자연어 처리 분야에서는 트랜스포머 기반 모델을 사용하는 방법으로 크게 BERT, GPT 계열이 있다

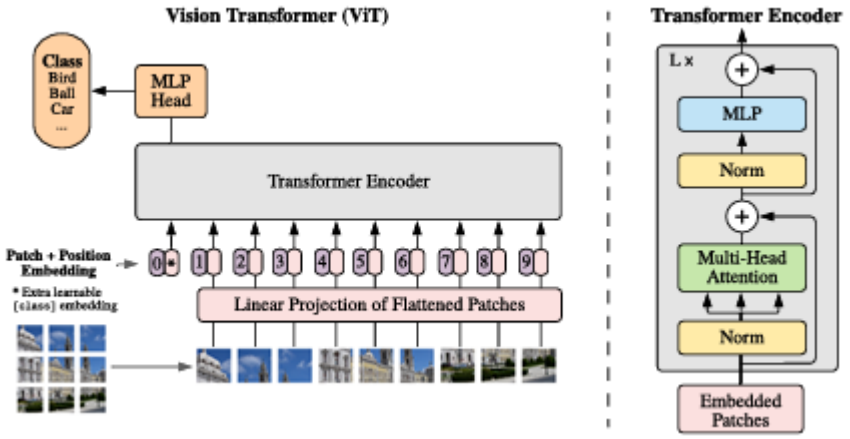
Vision 분야에 Attention 기법을 직접적으로 적용해 본다면 각 픽셀과 모든 픽셀의 attention을 계산해야 해 실제 데이터에 적용할 수 없다. 그래서 어텐션을 하나의 쿼리 픽셀과 근방을 픽셀들과의 어텐션만 계산하고, 다양한 사이즈의 블록이나 같은 축에 대해서만 어텐션을 적용한다.

전체 이미지에 어텐션을 적용하기 힘든 경우에 2x2 이미지 패치를 이용해 어텐션을 적용해 계산량을 줄인 연구가 있다.

ViT는 사전 학습을 통해 트랜스포머 성능을 CNN 모델과 비슷하게 보여준 장점이 있다.

3. Method

기존의 트랜스포머 구조를 거의 따른다.



3.1 Vision Transformer(ViT)

이차원 이미지를 다루려면 원본 이미지의 가로, 세로 픽셀, 채널 개수를 flatten 작업을 해 이미지 패치의 가로, 세로 픽셀 곱과 reshape이후 나오는 이미지 패치 개수로 reshape 해 토큰화한다. 인풋 시퀀스는 D차원의 잠재 벡터를 가짐.

patch embedding: 이미지 패치를 D차원으로 매핑시키는 linear projection 과정

class embedding: BERT 의 클래스 토큰과 비슷하게 임베딩 패치에 클래스 임베딩 추가. 아웃풋에서 이미지 라벨 반환하는 역할

positional embedding: 포지션별 정보를 유지하기 위해 더해짐. 2D 로 학습했을 때 특별히 좋은 결과가 있지 않아 1차원으로 학습. 인코더의 인풋으로 활용

Inductive bias(학습할 때 주어진 사전 지식, 가정): CNN에 비해 이미지 특화된 inductive bias 가 없다. 셀프 어텐션 층은 글로벌적이지만, MLP 층에서는 지역적으로, 번역적으로 동

등함. CNN처럼 지역적 정보를 명시적으로 학습하지 않아 inductive bias가 부족하지만 대규모 데이터와 자유로운 관계 학습으로 더 유연한 특징을 학습할 수 있다.

Hybrid architecture: 기존 이미지 패치 대신 인풋 시퀀스를 CNN으로 초기 특징 추출.

3.2 Fine-Tuning and Higher Resolution

보통 대규모 데이터셋에 ViT를 사전 학습 시키고 더 작은 태스크에서 fine tuning 진행

작은 태스크에 적용하기 위해 prediction head를 없애고 feed forward layer로 변경. 기존 이미지 해상도보다 고해상도로 fine tuning 할 때 효과적.

고해상도 이미지를 사용할 때는 사전 학습에 사용해 패치 크기와 동일한 사이즈를 사용해 더 긴 시퀀스 길이 사용하고 2차원 interpolation 진행

4. Experiments

representation learning 능력 평가

4.1 Setup

dataset: 대규모 이미지 데이터셋을 사전 학습한 후 사전 학습한 모델을 벤치마크 태스크에 전이함.

model variants: BERT에서 사용된 구성을 기반으로 base, large, huge 모델 생성. 트랜스포머의 시퀀스 길이는 패치 사이즈에 반비례함.

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Training & Fine-tuning: optimizer로는 Adam 사용. weight decay=0.1, linear learning rate warmup and decay. SGD, 배치 사이즈 512

Metrics: few-shot accuracy, fine-tuning accuracy 사용

4.2 Comparison to State of the Art

기존의 SOTA 모델은 CNN, 비교할 모델은 ViT-H/14와 ViT-L/16

Big Transfer: large ResNet 사용해 지도 전이 학습

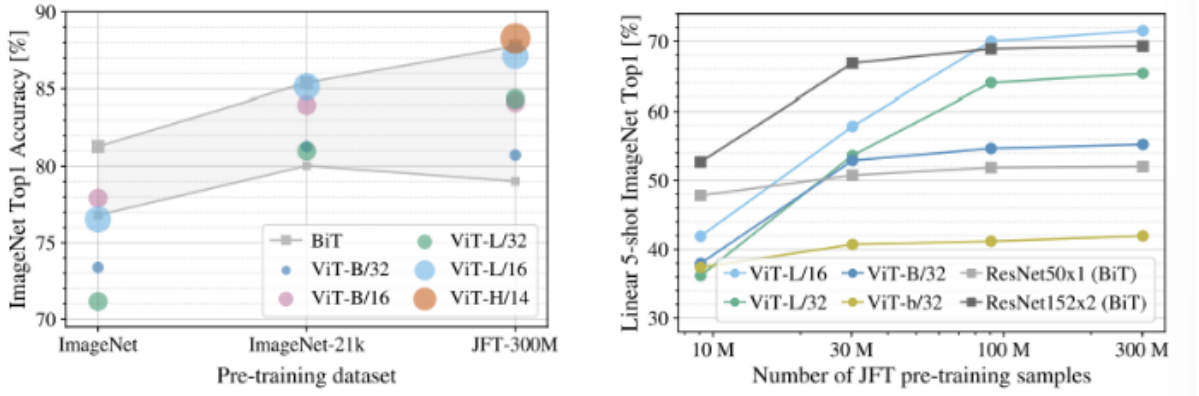
Noisy Student: large EfficientNet을 사용해 반지도 학습

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

→ 비슷한 성능에 압도적으로 적은 계산량

4.3 Pre-Training Data Requirement

ViT는 CNN보다 낮은 inductive bias를 가지기 때문에 더 많은 데이터가 필요.

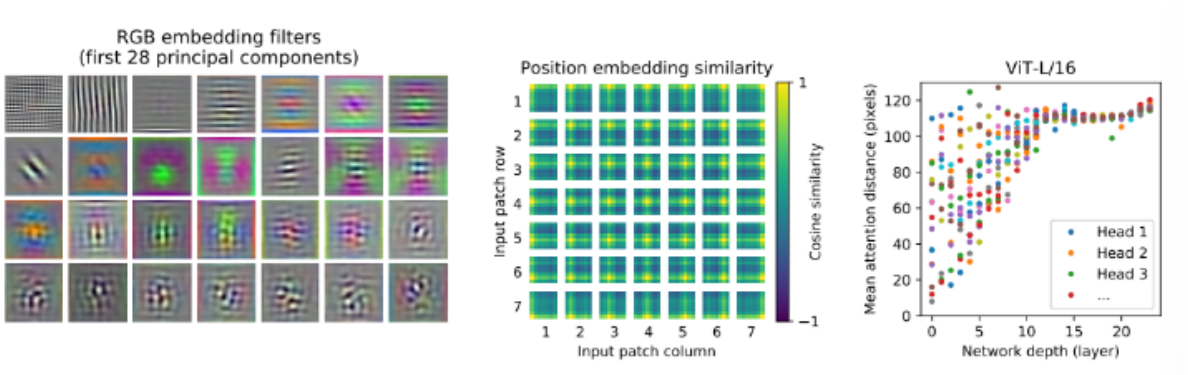


데이터 크기가 증가하는 방향으로 학습(10M, 30M, 100M, 300M)하면 점점 Linear fewshot 증가

4.4 Scaling Study

ViT 모델의 크기와 성능 사이 관계 분석. 더 큰 데이터셋과 더 많은 모델 파라미터를 사용할수록 성능이 좋아짐. 대규모 데이터셋에서 CNN보다 ViT가 더 좋은 성능을 보인다.

4.5 Inspecting Vision Transformer



1. ViT의 첫번째 layer는 flattened patch를 저차원 공간으로 투영함
2. 투영 후 positional embedding이 더해진 것에 대해 유사도 측정
3. 이미지 공간에서 평균 거리를 attention weight 기반으로 시각화

4.6 Self Supervision

트랜스포머는 NLP에 좋은 성능을 보이는데 뛰어난 scalability 뿐만이 아니라 대규모의 self-supervised pre-training 과정이 있었기 때문에 더 좋은 성능을 보임.

5. Conclusion

이미지 인식에 있어 트랜스포머를 적용함. 기존의 computer vision 에서 self-attention을 사용한 것과 달리 이미지 특화된 inductive bias를 초기 패치 단계에 넣지 않음. 대신 이미지를 패치의 시퀀스로 해석해 기본 트랜스포머 인코더로 처리. 특히 대규모 데이터셋에서 사전 학습할 때 좋은 성능을 보임.

발전 방향: ViT를 다른 CV분야(탐지, segmentation)에 적용. self-supervised pre-training 방법에 대한 더 깊은 연구 필요ㅇ