Paper Review 2

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

https://arxiv.org/abs/2010.11929

본 논문은 Vision 분야에서 활발하게 사용되지 않던 transformer architecture 를 적용한 ViT(Vision Transformer)에 관한 연구이다. 기존과 다르게 CNN 에 의존하지 않고 transformer 만을 이용하여 이미지 분류에 좋은 성능을 나타내는 ViT모델을 제안한다.

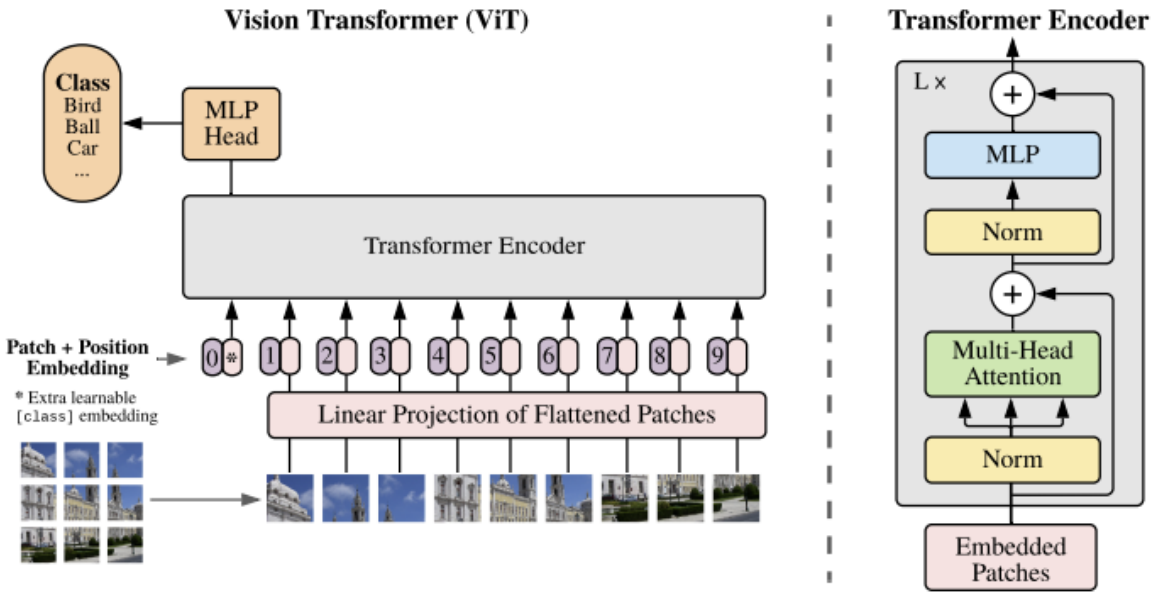
1. Introduction

Self-attention 기반 architecture(Transforemr)는 NLP분야에서는 활발하게 사용되었지만, Vision 분야에서는 convolution architecture가 지배적이었고, 큰 규모의 이미지 인식에서는 ResNet기반 architecture들이 SOTA모델이었다.

본 연구의 ViT는 이미지를 패치로 분할하고, 이 패치들을 linear 하게 임베딩하여 시퀀스를 생성한 후, 시퀀스를 transformer의 입력값으로 사용하였다. 이러한 방식으로 이미지 분류 모델을 지도학습하였다. 이는 ResNet과 큰 성능 차이를 보이지 않았다.

ViT가 ResNet과 다른 점은 inductive bias가 없다는 것이다. CNN 기반 ResNet은 translation equivariance와 locality 와 같은 가정이 존재하기 때문에, 데이터가 적을 때 성능이 높다. 하지만 large-scale dataset(14M-300M images)에서는 ViT가 SOTA 성능을 보인다.

2. Architecture of ViT



이미지를 패치로 분할하고, 이 패치들을 선형 임베딩하고 위치 임베딩을 더하여 시퀀스를 생성한 후, 시퀀스를 transformer encoder의 입력값으로 사용한다. 또한, 이미지 분류를 수행하기 위해 시퀀스에 classification token을 추가한다.

3. Method

H x W x C 형태의 이미지를 N x (P^2 * C)의 형태로 reshape 한다. 그 후, P^2 * C 차원의 이미지 패치를 D차원으로 매핑하는 linear projection을 수행하고, patch embeddings를 출력한다. class embeddings를 추가한 patch embeddings에 positional embedding를 더한 값이 transformer의 입력값이 된다.

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, \qquad \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D},\ \mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D} \tag{1}$$

$$\mathbf{z}'_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \qquad \ell = 1 \ldots L \tag{2}$$

$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \qquad \ell = 1 \ldots L \tag{3}$$

$$\mathbf{y} = \text{LN}(\mathbf{z}_L^0) \tag{4}$$

이러한 ViT를 큰 데이터셋에 pre-train 한 뒤, downstream tasks에 fine-tuning을 진행한다. 사전 학습할 때의 이미지 해상도(resolution)보다, 고해상도로 down-stream task에 fine-tuning을 하는 것이 효과적이다. 이 때, 해상도를 조정하고, 패치를 추출하는 과정이 ViT에서 유일하게 inductive bias가 들어가는 부분이다.

4. Experiments

다양한 사이즈의 데이터셋에 pre-train 한 결과, ViT가 낮은 pre-training cost로 SOTA를 기록하였다.

- Setup

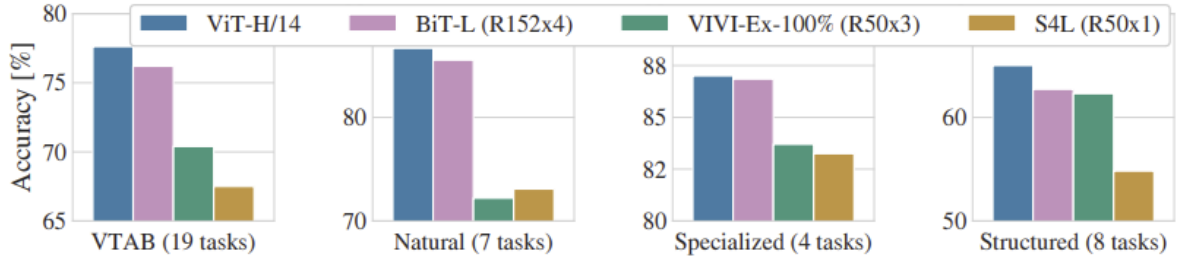
ViT-Base, ViT-Large, ViT-Huge 모델을 large-scale dataset (ILSVRC-2012 ImageNet dataset, ImageNet-21k, JFT) 에 pre-train 한 후, benchmark tasks(ImageNet, CIFAR-10/100, Oxford-IIIT Pets, Oxford Flowers-102)에 transfer 한다. 또한, 19-task 의 VTAB 분류도 평가하였다.

Training 과 Fine-tuning 과정에서는 각각 Adam과 SGDoptimizer를 사용하였다. 마지막으로, few-shot accuracy와 fine-tuning accuracy를 사용하여 downstream datasets의 결과를 기록하였다.

- Comparison to State Of The Art

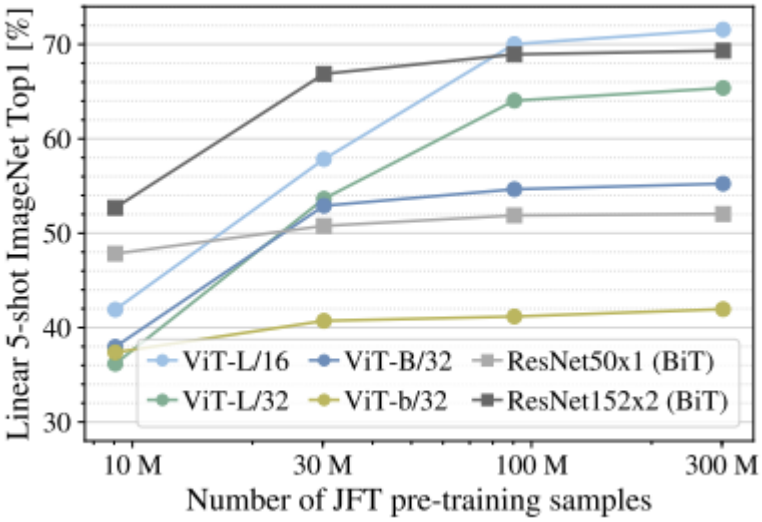
	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 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

각각의 데이터셋에서 JFT-300M에 pre-trained 한 ViT-H/14와 ViT-L/16이 BiT-L보다 성능이 좋고 연산량은 낮은 것을 알 수 있다.



VTAB(Natural, Specialized, Structured)에서도 ViT-H/14가 가장 좋은 성능을 나타내는 것을 알 수 있다.

- Pre-training data requirements



ViT-B/32는 ResNet50 9M subset일 때는 성능이 더 안 좋지만, 90M+에서는 성능이 더 좋아진다. 이는 ViT-L/16과 ResNet152도 마찬가지이다.

이를 통해, convolutional inductive bias 는 작은 데이터셋에서는 효과적이지만, 큰 데이터셋에서는 데이터에서 직접 패턴을 학습하는 것이 더 효과적이라는 것을 알 수 있다.

5. Conclusion

본 연구의 ViT는 기존과 다르게 NLP와 같은 방식으로 이미지를 패치들의 시퀀스로 나타낸 뒤, 이를 standard Tranformer encoder로 사용하였다. 이는 큰 데이터셋에서 pre-train을 할 때, 기존 CNN기반 ResNet 보다 낮은 cost로 높은 정확도를 기록하는 SOTA 성능을 보인다. 이러한 ViT는 classification 외의 detection과 segmentation과 같은 다른 비전 task에도 적용이 가능해 보인다.