Abstract

트랜스포머 구조가 자연어 처리 task 들에서 사실상 표준이 되는 동안, 비전에 이를 적용한 사례는 한정되어왔다. 비전 분야에서 어텐션은 Convolutional network 과 함께 적용되거나, Convolutional network 의 특정 요소를 대체하기 위해 사용되었기 때문이다.
✳ 이 논문에서는 이러한 **CNN 에 대한 의존이 필요하지 않고 순수 트랜스포머가 곧바로 이미지 패치들에 사용되고 이미지 분류에 잘 작동함**을 보여준다.

많은 양의 데이터에서 사전학습되고 여러 중간 사이즈나 작은 사이즈의 이미지 인식 데이터셋(ex. ImageNet, CIFAR-100)으로 전이 학습될 시에 Vision Transformer 는 기존의 convolution 기반의 SOTA 결과들에 비교해서 계산량은 적으면서 훌륭한 결과를 보여준다.

Introduction

Self-attention 기반의 구조들은 많이 알려졌듯이 자연어 처리분야에서 많이 사용되어 왔다. 특히 큰 텍스트 코퍼스에서 사전 학습하고 작은 task-specific dataset 에서 fine-tuning 하는 BERT 와 같은 방식이 우세하다.

하지만, 비전분야에서는 Convolutional 구조가 아직까지 우세하게 사용되고 있다. NLP 에서의 성공에 영감을 받아서 다양한 연구에서 CNN 같은 구조를 self-attention 과 결합하려고 시도해왔는데, 가장 최근의 모델들은 이론상으로는 효과적이었지만, 특수한 어텐션 패턴 사용 때문에 현대 하드웨어 가속기에서 효과적으로 **스케일링**이 불가능했다.

✳ NLP 에서의 트랜스포머 스케일링이 성공한 것에 영감을 받아, 이 논문에서는 standard transformer 를 최소한의 수정으로 직접 이미지에 적용하는 것에 대해 실험을 했다. 이를 위해, 이미지를 패치별로 쪼개고 (이미지 패치들은 NLP 에서의 token 과 같은 방식으로 다뤄지게 된다.), 이러한 패치들의 **linear embeddings sequence** 를 트랜스포머에 input 으로 넣었다.
이 모델을 supervised 방식으로 이미지 분류에 학습을 시켜 실험하였다.

ImageNet 과 같은 중간 사이즈의 데이터셋에서 학습했을 때, 모델은 ResNet 보다 약간 낮은 수치의 정확도를 보여줬는데, 이는 트랜스포머가 CNN 에 내재되어 있는 inductive

biases(translation equivariance and locality)가 부족함을 의미한다. 이 때문에 중간 사이즈의 데이터셋은 이 모델을 학습시키기에 충분하지가 않음을 알 수 있다.

하지만, 모델이 더 큰 데이터셋에서 학습되었을 때는 다른 양상을 보여주는데, 큰 스케일의 학습이 inductive bias 를 이겨버리는 것을 알 수 있었다. ViT 가 충분한 스케일의 데이터셋에서 사전학습되고 task 로 전이 학습될 때 훌륭한 결과를 보여줬기 때문이다.

특히 가장 잘 작동한 모델은 ImageNet 에서 88.55%, ImageNet-ReaL 에서 90.72%의 정확성을 보여주어 SOTA 와 비슷하거나 더 나은 결과를 보여주었다.

Method

모델 디자인에서 ViT 는 **오리지날 트랜스포머와 가능한 비슷하게** 디자인했는데, 이러한 심플한 setup 을 통해 스케일링이 가능하고 효율적인 실행이 가능했다.

1) Vision Transformer(ViT)

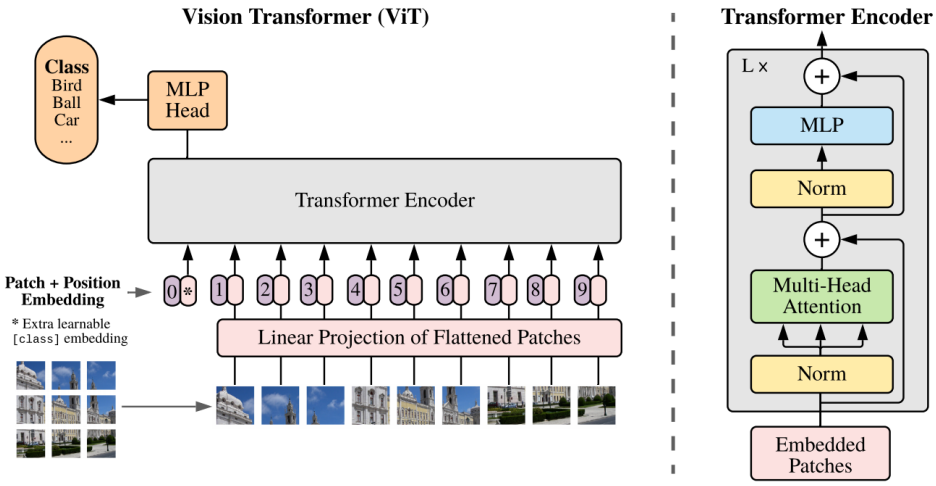

Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).

전체 모델의 구조는 위와 같은데, 이미지를 여러 패치들로 쪼개고 Linear projection 한 것과 위치 embedding(BERT 에서도 position embedding 을 사용했으니 여기서도 사용한 것 같다.)을 Transformer Encoder 에 넣어준다.

Standard Transformer 는 token embeddings 의 1 차원의 sequence 를 input 으로 받게 된다.
2 차원의 이미지(사실 채널까지 고려하면 3 차원이라고 해야 맞긴 하다)를 다루기 위해서, image x를 flattened 2 차원 패치들의 sequence 로 모양을 바꿔준다.

$₩mathbf {x} ₩in ₩mathbb {R}^{H ₩times W ₩times C}$
👉 $₩mathbf{x}_{p} ₩in ₩mathbb {R}^{N ₩times₩left(P^{2} ₩cdot C₩right)}$

여기서 (P, P)는 각 이미지 패치의 해상도가 되며 $N={HW}/{P^2}$는 패치의 수가 된다.
(정말 이미지들을 패치별로 자르고 납작하게 1 차원 벡터로 만들어서 이어줬다고 생각하면 된다.)

트랜스포머는 일정한 letent vector 사이즈 D를 모든 레이어에 걸쳐 사용하는데, 이를 통해 패치들을 flatten 시키고 D 차원으로 매핑시키게 된다. (이 projection 의 output 을 patch embedding 이라 한다.)

BERT 의 [class] 토큰과 비슷하게, ViT 의 저자들은 학습가능한 임베딩을 임베딩 된 패치들($₩mathbf {z}_{0}^{0}=₩mathbf {x}_{₩text {class }}$)의 sequence 에 추가하는데, 트랜스포머 encoder($₩mathbf {z}_{L}^{0}$)의 output 에서의 state 가 image representation y로 역할을 하게 된다.
Classification head 는 위의 그림에서도 확인할 수 있듯이 MLP 에 의해서 수행되며 사전학습때는 one hidden layer, fine-tuning 시에는 single linear layer 로 수행된다.

Position embedding 은 patch embedding 에 위치 정보를 위해 추가되는데, standard learnable 1D position embedding 을 사용했다. (2D 의 다른 방식도 써봤지만 그리 성능이 향상되지 않았다고 한다.)

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

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

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

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

위의 figure 1 에서 확인할 수 있듯이 Encoder 는 L 개로 되어있으므로 위의 식에서 l은 1 부터 L 까지로 되며 iteratively 계산되어 최종적으로 representation y가 나온다.

<Hybrid Architecture> : CNN feature map 을 적용한 방법

그냥 이미지 패치들을 사용하는 것 대신 Input sequence 는 CNN 의 feature map 으로 사용할 수도 있는데, 이 하이브리드 모델에서, 패치 임베딩 projection E는 CNN feature map 으로부터 뽑아낸 패치들에 적용된다. 특별한 경우로는 패치들이 spatial size 1X1 를 가질 수 있는데 이는 input sequence 가 feature map 의 spatial dimension 을 flatten 시키고 트랜스포머 차원으로 projecting 시킴으로써 나온 것을 의미한다. classification input embedding 과 position embedding 은 위와 같은 방법으로 추가된다.

2. Fine-Tuning And Higher Resolution

전형적으로 ViT 는 큰 데이터셋에서 사전학습되고 더 작은 downstream task 에서 fine-tuning 된다.
이를 위해서 저자들은 사전 학습된 prediction head 를 없애고 0 으로 초기화된 DXK 차원의 feedforward layer 을 붙인다.(K 는 downstream task 의 클래스 수)

사전학습 때보다 고해상도 데이터셋에서 fine-tuning 하는 게 어떨 때는 이득이라고 한다.(왜지???)
고해상도 이미지를 넣을 때, 패치 사이즈는 동일하게 하는데 이는 더 큰 효과적인 sequence length 를 만들어낸다.(더 고해상 도니까 패치수가 많아지고 더 길이도 길어짐) ViT 는 임의의 sequence 길이를 다룰 수 있긴 하지만, 사전 학습된 position embedding 은 의미가 없어지기 때문에, 사전 학습된 position embedding 의 **2 차원 interpolation** 을 오리지널 이미지에서의 위치에 따라 적용한다.

Experiments

실험을 위해서 ResNet, ViT 그리고 hybrid(CNN feature map 적용 모델)의 Representation learning capability 를 평가했다. 다양한 크기의 데이터셋으로 사전학습을 진행하고 많은 밴치마크 task 에 대해서 평가를 했는데. 사전학습 계산비용에 대해서 ViT 가 매우 효율적이었으며 대부분의 벤치마크에서 낮은 사전학습 비용으로 SOTA 수준의 성능을 달성했다.
마지막으로, self-supervision 을 사용해서도 작은 실험을 진행했는데 이에 대해서는 후속 연구에서 다룰 예정이다.

SETUP

<Datasets>

- Pre-training
모델의 스케일 능력을 실험하기 위해 ImageNet 데이터셋, ImageNet-21K(21K 개의 클래스와 1400 만 개의 이미지)과 JFT(18k 개의 클래스와 303M 의 고해상도 이미지)를 사용했다.

- Transfer Learning

전이 학습에 사용된 데이터셋으로는 ImageNet, CIFAR 10/100, Oxford-IIIT Pets 등이 있다.

<Model Variants>

BERT 에 사용된 구성을 기본으로 실험을 했는데 결과에 나온 B 는 Base, L 은 Large 그리고 H 는 Huge 를 뜻한다.

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

Table 1: Details of Vision Transformer model variants.

<Training&Fine-tuning>

- Pre-training

사전학습에 사용된 파라미터로는 Adam optimizer 의 $₩beta_1 =0.9$, $₩beta_2 =0.999$, 그리고 배치 사이즈는 4096 으로 실험하였다.

(+ weight decay parameter 는 0.1 로 설정)

- Fine-tuning

Stochastic Gradient 를 momentum 과 함께 사용했는데 여기서 배치사이즈는 512 로 설정하였다.

+ linear learning rate warmup 과 decay 도 사용하였다.

고해상도 이미지 사용으로는 ViT-L/16 에는 512, ViT-H/14 에는 518 을 사용하였다.

Comparison to SOTA

그럼 기존의 모델보다 얼마나 잘되는지 한 번 수치로 확인해보자.

	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

Table 2: Comparison with state of the art on popular image classification benchmarks. We report mean and standard deviation of the accuracies, averaged over three fine-tuning runs. Vision Transformer models pre-trained on the JFT-300M dataset outperform ResNet-based baselines on all datasets, while taking substantially less computational resources to pre-train. ViT pre-trained on the smaller public ImageNet-21k dataset performs well too. *Slightly improved 88.5% result reported in Touvron et al. (2020).

ViT-L/16 을 보면 모든 데이터셋에서 BiT-L 과 비슷하거나 더 나은 수치를 보여준다. 중요한 것은 TPUv3 로 학습한 비용을 보면 월등히 적다는 것이다.
Huge 모델인 ViT-H/14 은 기존의 모델보다 훨씬 더 나은 성능을 보여줌을 확연히 알 수 있다.

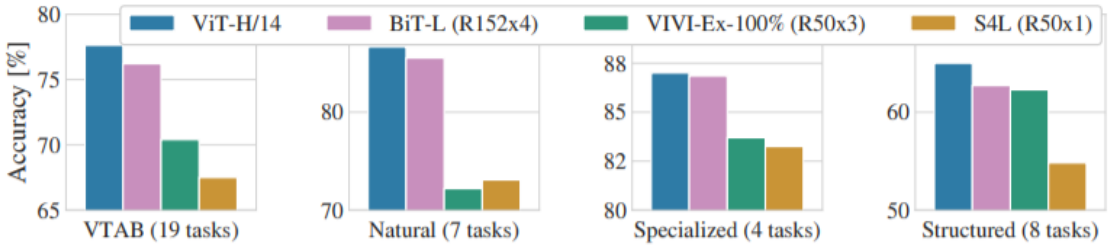

Figure 2: Breakdown of VTAB performance in *Natural*, *Specialized*, and *Structured* task groups.

위는 VTAB task 에 대해 기존의 SOTA 와 비교한 그래프인데 ViT-H/14 가 모든 task 에서 가장 뛰어난 정확도를 보여주는 것을 알 수 있다.

Pre-training Data Requirements

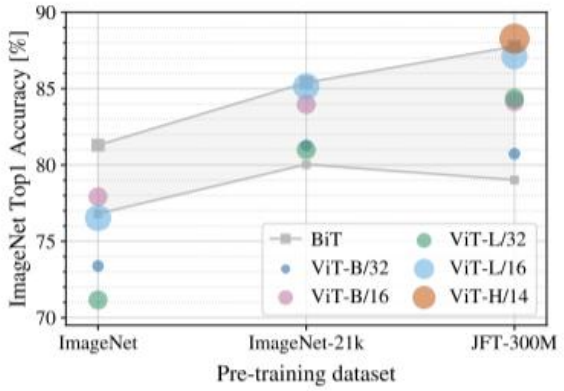

Figure 3: Transfer to ImageNet. While large ViT models perform worse than BiT ResNets (shaded area) when pre-trained on small datasets, they shine when pre-trained on larger datasets. Similarly, larger ViT variants overtake smaller ones as the dataset grows.

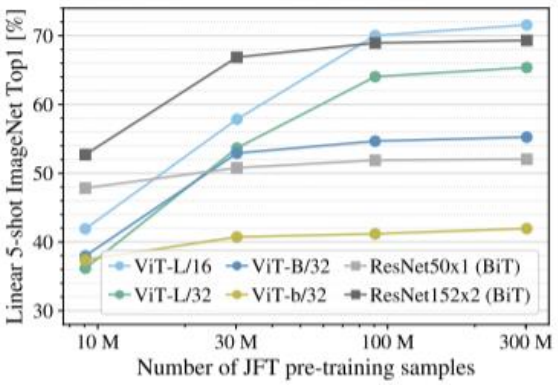

Figure 4: Linear few-shot evaluation on ImageNet versus pre-training size. ResNets perform better with smaller pre-training datasets but plateau sooner than ViT, which performs better with larger pre-training. ViT-b is ViT-B with all hidden dimensions halved.

Introduction 에서 언급했듯이 (아래 인용 참고)

ImageNet 과 같은 중간 사이즈의 데이터셋에서 학습했을 때, 모델은 ResNet 보다 약간 낮은 수치의 정확도를 보여줬는데, 이는 트랜스포머가 CNN 에 내재되어 있는 inductive biases(translation equivariance and locality)가 부족함을 의미한다. 이 때문에 중간 사이즈의 데이터셋은 이 모델을 학습시키기에 충분하지가 않음을 알 수 있다.

ViT 는 큰 데이터셋에서 사전 학습했을 때 충분히 학습이 되어 잘 작동한다고 했는데 위의 그래프가 이를 보여준다.

오른쪽 JFT-300M 데이터셋(큰 데이터셋)으로 갈수록 성능이 올라가는 것을 알 수 있다.
즉 데이터셋의 크기가 ViT 사전학습에 큰 영향을 끼치는 것을 다시 한번 알 수 있다.

Scaling Study

또 Introduction 에서 언급했듯이 트랜스포머를 비전에 적용한 기존의 연구들은 스케일링이 불가능했는데 ViT 에서는 어떻게 스케일링이 가능한지에 대해 측정한 그림이 다음과 같다.

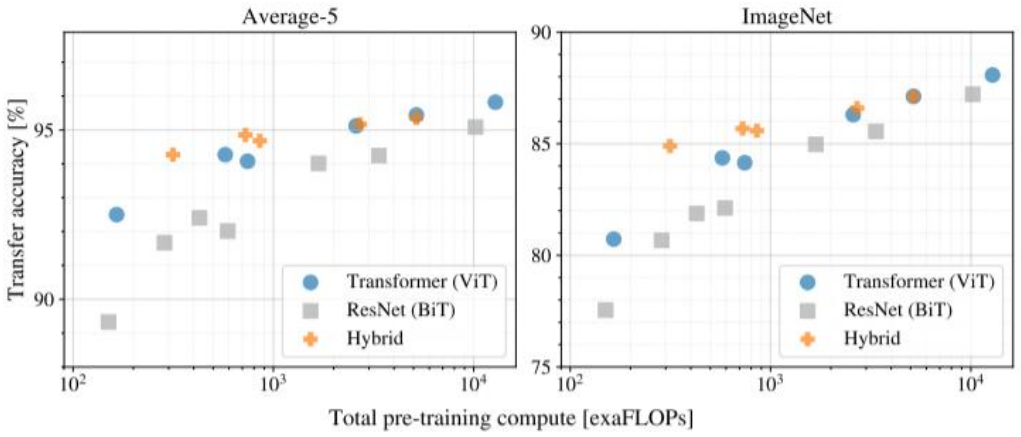

Figure 5: Performance versus cost for different architectures: Vision Transformers, ResNets, and hybrids. Vision Transformers generally outperform ResNets with the same computational budget. Hybrids improve upon pure Transformers for smaller model sizes, but the gap vanishes for larger models.

위의 그림을 통해 세 가지를 알 수 있는데,

1. ViT 는 ResNet 보다 동일한 성능을 내기 위해 반 정도의 컴퓨팅이 필요하다는 것

2. CNN 의 feature map 을 이용한 하이브리드 모델은 적은 computing cost 에서는 ViT 를 능가하지만 cost 를 늘리게 되면 큰 차이가 없어진다는 것.

3. ViT 는 "saturate" 되지 않으며 스케일링이 가능하다.

Inspecting Vision Transformer

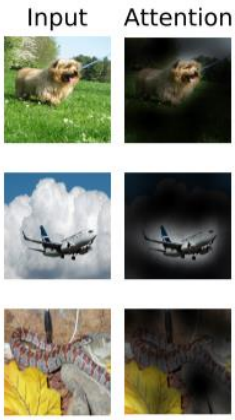

Figure 6: Representative examples of attention from the output token to the input space. See Appendix D.6 for details.

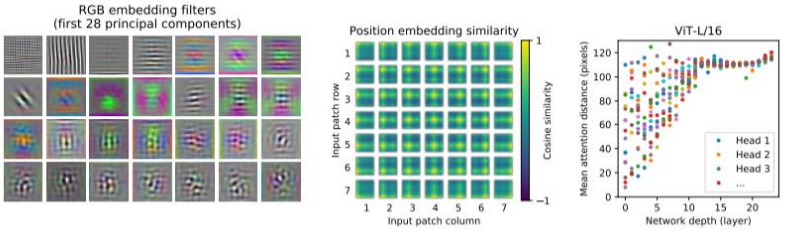

Figure 7: **Left:** Filters of the initial linear embedding of RGB values of ViT-L/32. **Center:** Similarity of position embeddings of ViT-L/32. Tiles show the cosine similarity between the position embedding of the patch with the indicated row and column and the position embeddings of all other patches. **Right:** Size of attended area by head and network depth. Each dot shows the mean attention distance across images for one of 16 heads at one layer. See Appendix D.6 for details.

다음은 ViT 가 어떻게 이미지를 처리하는지에 대해 분석한 내용이다.

주목할 점은 ViT 는 가장 하위의 layer 에서도 전체 이미지에 대한 정보를 통합할 수 있다는 점이다. (오른쪽 그림 참고)
낮은 Network depth 에서도 attention 을 통해 global 하게 정보를 사용할 수 있다는 점을 알 수 있다.(..!)
또한 depth 가 증가함에 따라 attention distance 도 증가함을 알 수 있다.

Self-supervision

후속 연구로 언급했던 self-supervision 에 대한 내용도 또 나왔는데, Self-supervised 사전학습을 한 ViT-B/16 모델은 ImageNet 에서 79.9%의 정확도를 보였지만 supervised 사전학습을 했을 때보다는 4% 정도 하락한 성능이라고 한다.

자세한 건 Appendix 참고.

Conclusion

이 논문에서는 이미지 인식에 기존 연구와 달리 직접적으로 트랜스포머를 사용한 방법을 제안하였다.
Convolution 기반의 모델은 inductive bias 를 사용하는 것과는 달리 대규모 데이터셋으로 사전학습을 통해 놀라운 성능을 보여주었다. 또한 스케일링 측면에서도 우수함을 확인할 수 있었다.

하지만 많은 challenge 도 남아있는데,

1. 인식 외에 detection, segmentation 과 같은 task 에 적용하는 것

2. 사전학습 방법에 대해 좀 더 연구하는 것(self-supervision 과 연결)

3. 위에서도 보았듯이 ViT 가 saturate 되지 않았는데 이를 통해 더 확장시켜 더 나은 성능을 확인할 수도 있는 것

이 세 가지가 있다.