

- Introduction.

Transformer의 성공 이후 CNN과 attention을 조합하는 시도가

많았지만 각 model의 특수한 구조 때문에 model의 효율은 좋지 않지만

SOTA는 여전히 Percept이다.

다른 transformer의 구조를 가지 바꾸지 않고, 이미지를 patch 단위로

token화 했고, ResNet의 구조를 성능을 냈다.

Transformer inductive bias가 없기 때문에 충분히 많은 dataset이라는 정규화가 되지 않아.

훈련이 잘되지 않았지만, JFT-300M같이 큰 모델에서는 inductive bias 훈련되는 것으로

나타났고, 인력한 성능을 보였다.

- Related Work.

원래 NLP의 transformer는 큰 corpora의 의해 학습된다.

Naive transformer의 image의 적용은 각 pixel이 모든 다른 pixel과의

attention을 계산해야 하는데 계산량이 증가 때문에 전처리 이미지에서 적용되기 힘들다.

다른 image processing with 많은 근사가 시도되었다.

1. local neighbor only만 self-attention을 수행하는 방법.

2. global-self-attention의 적용을 수 있도록 scalable한 근사 방법을 쓰는 것

3. 여러 block 크기를 쓰는 법

등이 있지만 hardware accelerator의 적용하기 복잡하다.

- Method

model을 디자인 할 때, 선택한 original transformer를 기반화 했음.

내 학장은 transformer를 쉽게 적용할 수 있다.

* ViT (Vision transformer)

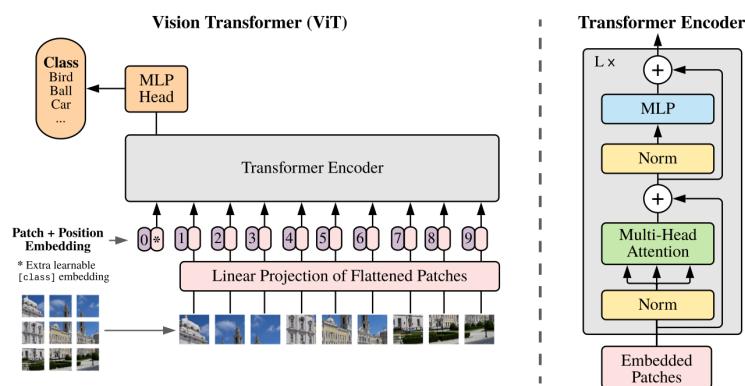


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable “classification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

ViT transformer는 embedding 된 ID token을 input으로 넣음.

2D를 1D로, $(P, P) \rightarrow$ patch size로 ∞ . image $x \in \mathbb{R}^{H \times W \times C} \equiv x_p \in \mathbb{R}^{N \times (P \times C)}$ resize 가능. $N = H \times W / P^2$

N 은 transformer의 input sequence로 되는 patch의 수 (sequence length)다.

Transformer는 D 차원 D dimension을 사용해 input의 patch를 $D \times D$ mapping (trainable linear projection) 한다.

↳ patch embedding

Bert의 [class] token과 유사하게, embedded patch sequence의 learnable embedding을 알 수가 있다.

Transformer encoder의 출력 (\mathbb{R}^d)의 output state는 image representation y 로 쓰인다.

pre-training과 fine-tuning 때마다 \mathbb{R}^d 의 classification head를 붙인다.

pre-training은 1 hidden layer MLP인 fine-tuning은 single linear layer인 classification head를 적용한다.

Image patch의 위치 정보를 모아 각각 PE를 patch embedding의 structure.

2D PE의 성능이 훨씬 느끼지 못하기 때문에 standard learnable 1D PE를 사용했다.

Transformer encoder는 multi-head self-attention과 MLP로 이루어진다.

Layernorm과 residual은 모든 block에서 쓰였으며, MLP는 GELU가 사용된다.

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

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

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

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

1) Inductive bias

Vision transformer는 CNN의 비슷한 inductive bias를 가졌다.

CNN은 local, translation equivariance는 모든 model의 각 layer에 적용된다.

↳ image에 대한 동일한 적용

ViT와 같은 MLP는 local, translation equivariance는 self-attention은 global이다.

1D PE는 사용하지 않지만 image cutting 방식으로, fine-tuning 때 PE 자체를 가끔 쓰인다.

1) 이전에도, PE가 처음 initialization 시에는 patch의 2D image의 위치 정보가 없고

patch 간 관계를 허용하지 않도록 한다.

2) Hybrid Architecture

Raw Image patch 대신, CNN의 feature map을 사용할 수 있다.

hybrid model은 patch embedding projection E는 CNN feature로부터 뽑아낸 patch를 적용할 수 있다.

특수 경우로, patch는 입력 sequence를 feature map의 spatial dimension을 flatten하고,

transformer dimension으로 projection을 1x1 spatial size로 풀어낸다.

= fine-tuning and hyper-resolution

일반적으로 pre-train은 큰 dataset(1M, downstream)에서 fine-tuning은 작은 dataset을 쓴다.

only then,这样才能 pre-trained prediction head를 제작함. Zero-initialized된 $D \times k$ FFM을 넣인다.

Resolution이 큰 image를处理할 때, patch size는 같은데 sequence length는 길어진다

ViT는 global sequence를 다룰 수 있지만 pre-trained 된 PE는 global PE는 아니기 때문이다.

PE의 1D interpolation을 수행한다

↳ ViT의 주요한 핵심은 inductive bias가 반영되는 과정

- Experiment

ResNet, ViT, Hybrid의 성능을 비교했다.

각 model의 data를 이용해 각각의 다양한 size의 benchmark에 평가를 진행했다.

pre-training의 computational cost를 고려했을 때, ViT는 다른 pre-training cost로

대부분의 benchmark에서 SOTA를 달성했다.

또한, self-supervision의 대안 혹은 실험을 통해, self-supervision이 이후 ViT 연구에 유용하다는 것을 보인다.

* Setup

1) Datasets

model의 scalability를 위한 ImageNet-1k(1300M), ImageNet-21k(14000M), JFT-300M(18k class)

를 사용했다. downstream tasks의 대상은 pre-training dataset을 사용한다.

이 dataset들이 자체 훈련된 모델로 여러 benchmarks에 대해 transfer된다.

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.

2) Model Variants

ViT의 configuration은 table 1과 같다.

Baseline CNN으로는 ResNet(BN, LReLU, Group Normalization, standardized conv)을 사용한다. (ResNet(BiT))

Hybrid로는, 중간 feature map을 디코딩 pixel patch로 ViT로 준다.

다른 sequence length로 실험하기 위해, 1) ResNet50의 stage 4의 흡수, 2) stage 4를 제거하고, stage 3의

동일한 수의 layer를 넣었지만, stage 3의 출력은 같았다

2)는 sequence length가 448이 되어 ViT model의 성능에 영향을 미친다.

3) Training & fine-tuning

ResNet은 포함된 모든 model을 (0.9, 0.999) Adam와 4096 bs, high weighted decay 0.1을 적용한다.

fine-tuning 때 bs 512, SGD with momentum을 사용.

↳ Appendix B

4) Metrics

few-shot 또는 fine-tuning을 통한 downstream 결과를 보인다.

Fine-tuning accuracy는 fine-tuning을 몇 번에 진행 흐름을 따른다.

Few-shot accuracy는 훈련 이미지의 subset의 representation을 $\{-1, 1\}^k$ target vector로

mapping 시에 regularized least-square regression을 사용함으로써 얻는다.

각각 fine-tuning을 위한 cost가 매우 큰 경우 few-shot을 선택

→ Comparison to SOTA

$\text{ViT-L/16} \approx \text{ViT-H/14}$ ↳ Big Transfer, Noisy student ↳ 사용자

	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).

⇒ SOTA보다 둘째 줄입니다

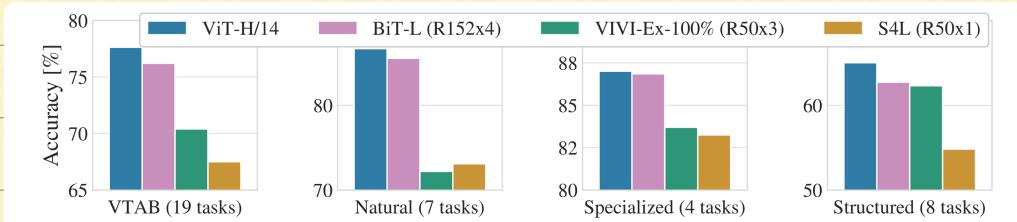


Figure 2: Breakdown of VTAB performance in *Natural*, *Specialized*, and *Structured* task groups.

* Pre-training data requirements

ViT는 JFT처럼 큰 dataset을 than pre-training 했을 때, 동작을 잘 한다.

ResNet와 ViT의 inductive bias가 다른데 이를 통해 ViT의 실험은 진행했습니다.

둘째로, ViT의 pre-training size를 늘리면서 성능이 향상되었습니다. ImageNet, ImageNet-21k, JFT

작은 dataset은 weight decay, dropout, label smoothing, regularizing을 적용했습니다.

ImageNet의 작은 dataset ↓

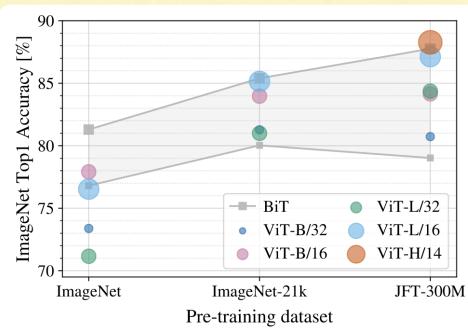


Figure 3: Transfer to ImageNet. While large ViT models perform worse than BiT ResNets (shaded area) when pre-trained on small datasets, they shine when pre-trained on larger datasets. Similarly, larger ViT variants overtake smaller ones as the dataset grows.

pre-training의 작은 dataset에서 이전일 경우, ViT-Large가 더 뛰어나

ViT-base보다 성능이 낫습니다. (ImageNet)

ImageNet-21k의 경우 비슷한 성능입니다.

JFT-300M의 경우 큰 ViT의 성능이 큰 이득이 있습니다.

두 번째로, JFT-300M을 random으로 뽑은 9M, 36M, 96M의 subset을

적용하는 regularization 때문. 같은 hyper-param으로 훈련시켰다.

이 방법으로 model의 훈련 속도가 적은 토성은 유익하게 활용함.

Best val accuracy early stopping을 수행합니다.

그림에 full-finetuning 했을 때 few-shot linear acc은 얼마나 ↓

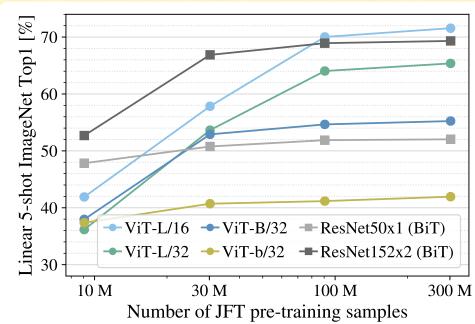


Figure 4: Linear few-shot evaluation on ImageNet versus pre-training size. ResNets perform better with smaller pre-training datasets but plateau sooner than ViT, which performs better with larger pre-training. ViT-b is ViT-B with all hidden dimensions halved.

ViT는 ResNet보다 더 작은 dataset에서 비슷한 computational cost로 overfitting이

작은 small dataset에서 convex inductive bias가 유용한데,

large dataset에서 dataset의真正 relevant pattern을 학습하는 것이 좋다.

전반적으로, ViT는 low data transfer를 유리한 것으로 보인다.

* Scaling study

각각의 model 크기로 조절되었고, JFT의 transfer 성능을 비교해보자.

pre-training cost vs transfer performance를 비교해보자.

model은 ResNet의 경우 72U ($R_{50} \times 1$, $R_{50} \times 2$, $R_{101} \times 1$, $R_{152} \times 1$, $R_{152} \times 2$ for 7 epoch CL

$R_{152} \times 1$, $R_{200} \times 3$ for 14 epoch)를 사용하고,

ViT는 62U ($B_{1/2}$, $B_{1/16}$, $L_{1/16}$, $L_{1/32}$ for 7 epoch CL

$L_{1/16}$, $H_{1/16}$ for 14 epoch)를 사용하고 CL.

Hybrid의 경우 52U ($R_{50} + ViT-B_{1/32}$, $B_{1/16}$, $L_{1/32}$, $L_{1/16}$ for 7 epoch CL

$R_{50} + ViT-L_{1/16}$ for 14 epoch)를 사용한다.

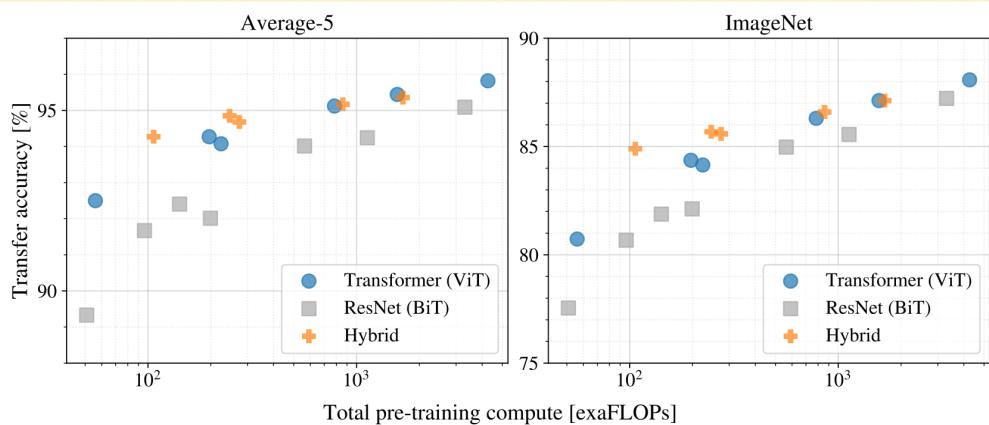


Figure 5: Performance versus pre-training compute for different architectures: Vision Transformers, ResNets, and hybrids. Vision Transformers generally outperform ResNets with the same computational budget. Hybrids improve upon pure Transformers for smaller model sizes, but the gap vanishes for larger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

→ detail은 Appendix!

→ 전기학적으론 ViT와 ResNet의 performance / cost trade-off가 명확해.

⇒ hybrid는 small computational budget 하에서 ViT보다 더 낮은 성능을 보여.

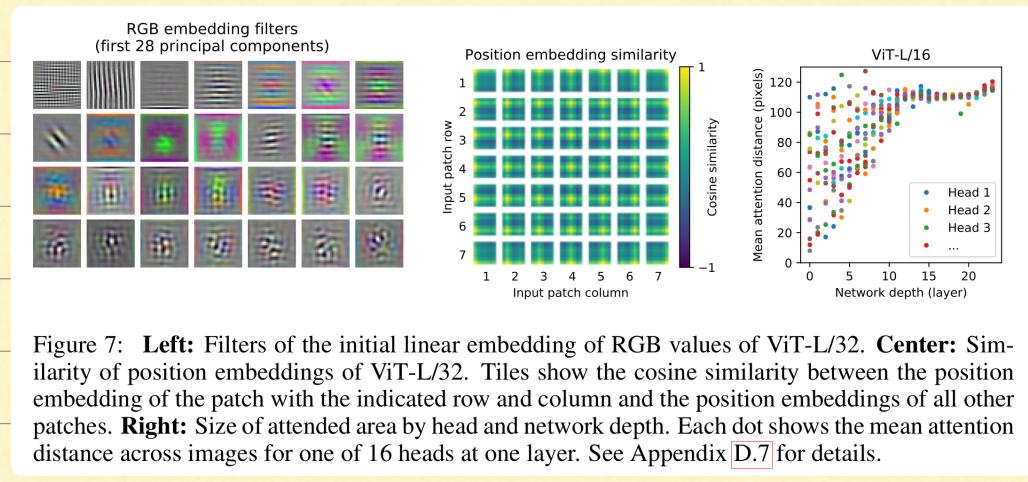
large model은 ViT와 비슷한 성능을 보여.

⇒ m-size에서 convolutional local feature 처리를 할 때는 ViT가 좋다.

⇒ ViT는 시도한 실험 대부분은 saturate되는 양이지만 초기 학습 가능성이 있다.

* Inspecting ViT

ViT의 image data process 과정 분석을 위해, internal representation의 구조를 살펴보자.



ViT의 첫 layer는 eq 1.과 같이 image를 low dimension의 flatten patch로 projection한다.

Fig 7의 왼쪽은 학습된 embedding filter로 first 28 principal components를 나타낸다.

\hookrightarrow patch의 fine structure의 low-dimension representation의 basis function을 나타낸다.

Projection 이후 learned embedding patch는 patch representation을 대체한다

Fig 7의 중앙은 model의 PE의 유사성으로 image-level 차이를 네트워크는 모든 이미지에

가까운 patch의 PE는 네트워크.

ViT의 self-attention은 전개 이미지에서의 정보를 통합할 수 있다.

기자는 network가 얼마나 이 기능을 사용하는지 살펴본다.

특히 attention weight를 기준으로, 어느 정보가 통합되는지 image space의 영역리를 살펴보았다. (Fig 7)

Attention distance는 current receptive field와 비슷하다

lowest (layer 0)에는 이미 몇몇 head가 전개 이미지에 대한 attention 빠르고 통합된 information은 전개적으로 보여지는 경우가 많다.

localized attention은 Transformer과 ResNet을 적용하는 hybridism의 나ether로, 이는 초기 conv layer의 비슷한 역할을 할 수 있음을 시사한다

high attention distance는 depth가 끝에 있는 노드를 늘어나는데, 모델은 classification과 관련이 있는

image regions의 attention 흐름 \Rightarrow Appendix D

* Self-supervision.

Transformer는 NLP에서 폭넓은 성능을 거두었는데 이는 scalability 때문이거나

large scale self-supervised pre-training로 낙타였다.

그러나 Bert와 같은 self-supervised를 위한 masked patch 방식은 이미 연구했다
서로 다른 훈련하는 것보다 훨씬 더监督된 self-supervised 모델의 안정성을.

- Conclusion

Image transformer를 바로 적용시키는 방식을 연구했다.

다른 inductive bias를 통해 architecture의 다른 image-specific을 더한 논리를 살펴보기

image patch를 sequence로 만들면서 기존 transformer의 적용했다.

간단하면서도 scalable 했고, large dataset으로 pre-training 했을 때 효과적이다.

VIT는 다른 SOTA의 비슷하거나 더 나은에서도 작은 pre-training cost를 가졌다.

하지만 detection이나 segmentation이 적용되고 self-supervision의 문제가 남아 있다.

\hookrightarrow self-supervised와 supervised 모델 큰 차이가 있는

- Appendix