VIT

pdf

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/052fd91c-a493-4dd7-ba8f-cad55ccfa29a/2010.11929v2.pdf

---

Summary

Inductive Bias - background

training에서 보지 못한 데이터에 대해서도 적절한 귀납적 추론이 가능하도록 하기 위해 모델이 가지고 있는 가정들의 집합. 즉, 모델이 목표 함수를 학습하고 훈련 데이터를 넘어 일반화하기 위해 만든 가정

▼ DNN의 기본적인 요소들의 inductive bias

- Fully connected
  입력 및 출력 element가 모두 연결되어 잇으므로 구조적으로 특별한 relational inductive bias를 가정하지 않음
- Convolutional
  CNN은 작은 크기의 kernel로 이미지를 지역적으로 보며, 동일한 kernel로 이미지 전체를 본다는 점에서 locality와 transitional invariance 특성을 가짐
- Recurrent
  RNN은 입력한 데이터들이 시간적 특성을 가지고 있다고 가정하므로 sequentiality와 temporal invariance 특성을 가짐

DNN: 입력층(input layer)과 출력층(output layer) 사이에 여러 개의 은닉층(hidden layer)들로 이뤄진 인공신경망(Artificial Neural Network, ANN)

- Transformer는 CNN및 RNN보다 상대적으로 inductive bias가 낮음
- ViT에서 MLP는 locality와 translation equivariance가 있지만, MSA는 global하기 때문에 CNN보다 image-specific bias가 낮음 (??)

  ▼ ViT에서는 모델에 아래 두 가지 방법을 사용하여 inductive bias의 주입을 시도함

  - Patch extraction
    image를 여러개의 Patch로 분할해서 순서가 있는 상태로 넣는 것 -(locality와 translation equivariance가 있다)
  - Resolution adjustment
    이미지의 크기(해상도)에 따라서 Patch의 크기는 동일하지만 생성되는 개수는 달라지는데 Fine-tuning을 할 때 다른 크기(해상도)의 이미지의 position embeding을 조정(inductive bias 조절)

▼ ViT 모델 구조

1. 이미지 $x \in R^{(H \times W \times C)}$ 가 있을 때, 이미지를 $(P \times P)$크기의 패치 $N(= H \times W / P^2)$개로 분할하여 패치 sequence $x_p \in R^{N \times (P^2 \cdot C)}$를 구축함
2. Trainable linear projection을 통해 x_p의 각 패치를 flatten한 벡터를 D차원으로 변환한 후, 이를 패치 임베딩으로 사용함
3. Learnable class 임베딩(*)과 패치 임베딩에 learnable position 임베딩을 더함
4. 임베딩을 vanilla Transformer encoder에 input으로 넣어 마지막 layer에서 class embedding에 대한 output인 image representation을 도출함
5. MLP에 image representation을 input으로 넣어 이미지의 class를 분류함

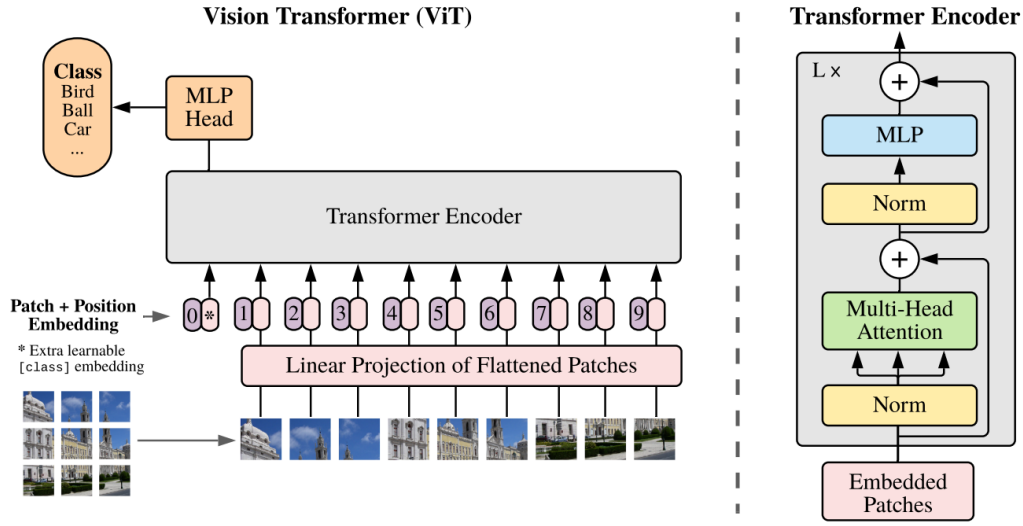


그림 1: 모델 개요. 이미지를 고정된 크기의 패치로 분할하고 각 패치를 선형으로 임베딩함. Patch에 Position EMbedding을 추가하고 벡터 시퀀스를 Transformer Encoder에 넣음. 분류를 수행하기 위해 학습 가능한 “classification Token”을 시퀀스에 추가하는 표준 접근 방식을 사용함.

▼ Positional Embedding

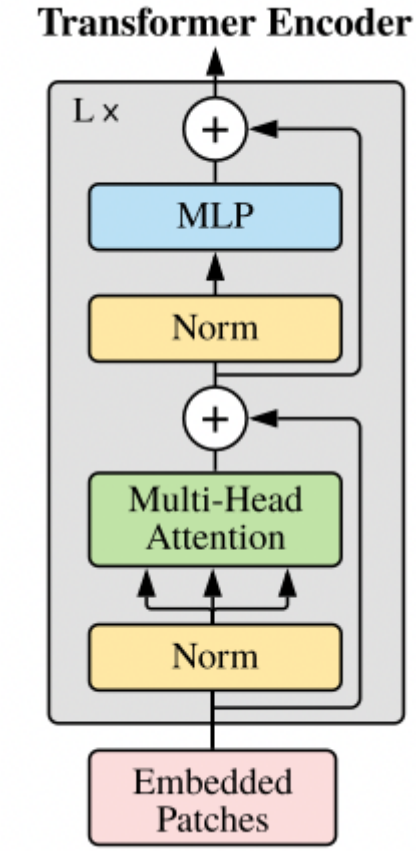
아래 4가지 position임베딩을 시도한 후, 최종적으로 가장 효과가 좋은 1D position 임베딩을 ViT에 사용함

1. No positional information
   패치의 임베딩만을 input으로 사용
2. 1-dimensional positional embedding
   왼쪽 위부터 오른쪽 아래까지 순서대로 패치를 보는 것
3. 2-dimensional
   x,y축의 좌표가 있는 positional embedding
4. Relative positional embeddings
   패치간의 상대적인 거리를 사용한 positional embedding

Pos. Emb.	Default/Stem	Every Layer	Every Layer-Shared
No Pos. Emb.	0.61382	N/A	N/A
1-D Pos. Emb.	0.64206	0.63964	0.64292
2-D Pos. Emb.	0.64001	0.64046	0.64022
Rel. Pos. Emb.	0.64032	N/A	N/A

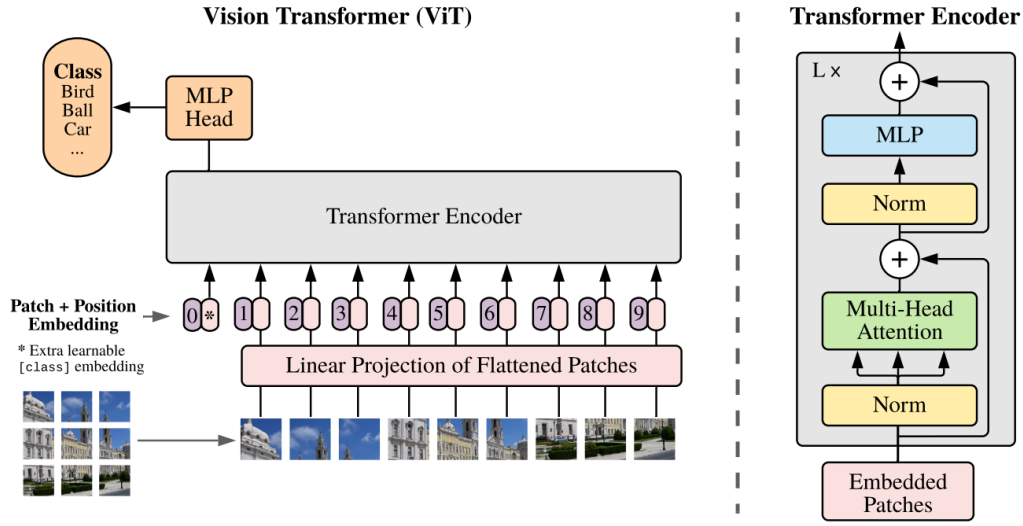
▼ Transformer Encoder

- ViT는 Multi-head Self Attention(MSA)와 MLP block으로 구성되어 있음
- MLP는 2개의 layer를 가지며, GELU activation function을 사용함
- 각 block의 앞에는 Layer Norm(LN)을 적용하고, 각 block의 뒤에는 residual connection을 적용함



▼ Hybrid Architecture

- ViT는 raw image가 아닌 CNN으로 추출한 raw image의 feature map을 활용하는 hybrid architecture로도 사용할 수 있음
- Feature map은 이미 raw image의 공간적 정보를 포함하고 있으므로 hybrid architecture는 패치 크기를 1×1로 설정해도 됨
- 1×1크기의 패치를 사용할 경우 feature map의 공간 차원을 flatten하여 각 벡터의 linear projection을 적용하면 됨



▼ Fine-tuning and Higher Resolution

- Large scale로 ViT를 pre-training한 후, 해당 모델을 downstream task에 fine-tuning 해 사용할 수 있음
- ViT를 fine-tuning할 때, ViT의 pre-trained prediction head를 zero-initialized feedforward layer로 대체함(Transformer Encoder는 그대로 사용하되 특정 태스크의 Output을 도출하기위한 MLP-Head는 풀고자 하는 Downstream task의 목적에 맞도록 학습하기 위해서 Zero-initialized feedforward layer로 대체함)
- ViT를 fine-tuning할 때, pre-training과 동일한 패치의 크기를 사용하기 때문에 고해상도의 이미지로 fine-tuning을 하면 sequence길이가 더 길어짐
- ViT는 가변적 길이의 패치들을 처리할 수 있지만, pre-trained position embedding은 의미가 사라지므로 pre-trained position embedding을 원본 이미지의 위치에 따라 2D interpolation하여 사용함(**Resolution adjustment**)

▼ Experiment

▼ Experimental Settings

Datasets

- ViT는 아래와 같이 class와 이미지의 개수가 다른 3개의 데이터셋을 기반으로 pre-train 됨
- 아래의 benchmark tasks를 downstream task로 하여 pre-trained ViT의 representation
  성능을 검증함

Pre-trained Dataset	# of Classes	# of Images
ImageNet-1k	1k	1.3M
ImageNet-21k	21k	14M
JFT	18k	303M (High resolution)

Model Variants

- ViT는 아래와 같이 총 3개의 volume에 대해 실험을 진행하였으며, 다양한 패치 크기에 대해 실험을 진행함
- Baseline CNN은 batch normalization layer를 group normalization으로 변경하고 standardized convolutional layer를 사용하여 transfer learning에 적합한 Big Transformer(BiT) 구조의 ResNet을 사용함 (???)

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.

▼ Comparison to SOTA

실험 결과

- 본 실험에서는 14 × 14패치 크기를 사용한 ViT-Huge와 16 ×16 패치 크기를 사용한 ViT-Large의 성능을 baseline과 비교함
- JFT 데이터셋에서 pre-training한 ViT-L/16 모델이 모든 downstream task에 대하여 BiT-L 모델(Baseline)보다 높은 성능을 도출함
- ViT-L/14 모델(14 ×14 패치)은 ViT-L/16 모델(16 ×16 패치)보다 향상된 성능을 도출하였으며, BiT-Large 모델보다 학습 시간이 훨씬 짧음

	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).

▼ 19-task VTAB classification suite를 아래와 같이 3개의 그룹으로 나누어 추가 실험을 진행

- Natural: tasks like Pets, CIFAR, etc
- Specialized: medical and satellite imagery
- Strictured: tasks that require geometric understanding like localization

- 전체 데이터뿐만 아니라 각 그룹에서도 ViT-H/14가 좋은 결과를 도출함

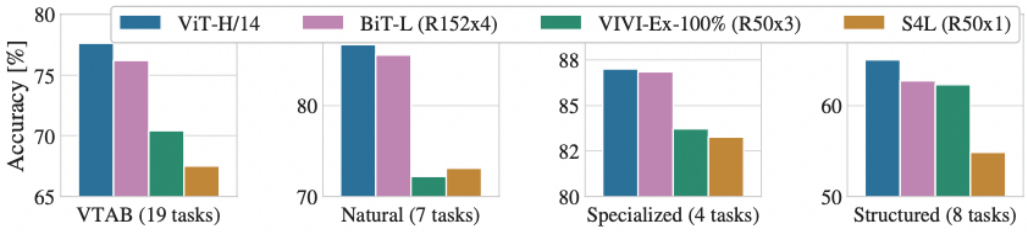


Figure 2: Breakdown of VTAB performance in *Natural*, *Specialized*, and *Structured* task groups.

▼ Pre-training Data Requirements

- 본 실험에서는 pre-training 데이터셋의 크기에 따른 fine-tuning성능을 확인함
- 각 데이터셋에 대하여 pre-training한 ViT를 ImageNet에 transfer learning한 정확도를 확인한 결과, 데이터가 클수록 ViT가 BiT(baseline CNN)보다 성능이 좋고 크기가 큰 ViT모델이 효과가 있었음 → pre-training 데이터가 커야 성능이 좋고, 데이터가 커질수록 ViT의 성능이 CNN보다 좋은 성능
- JFT를 각각 다른 크기로 랜덤 샘플링한 데이터셋을 활용(데이터셋의 크기가 달라질 때)하여 실험을 진행한 결과, 작은 데이터셋에서 CNN inductive bias가

효과가 있으나 큰 데이터셋에서는 데이터로부터 패턴을 학습하는 것만으로 충분함을 알 수 있음

- 즉, CNN이 작은 데이터셋에서는 inductive bias로 인하여 좋은 성능
  데이터가 많아지면 ViT가 좋은 성능

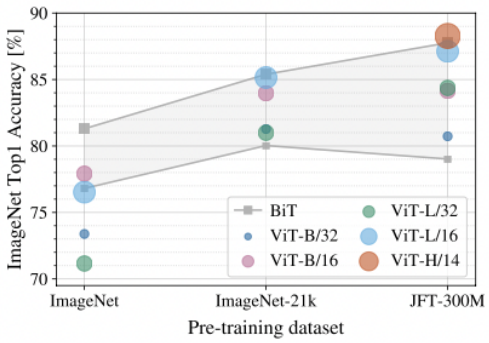


Figure 3: Transfer to ImageNet. While large ViT models perform worse than BiT ResNets (shaded area) when pre-trained on small datasets, they shine when pre-trained on larger datasets. Similarly, larger ViT variants overtake smaller ones as the dataset grows.

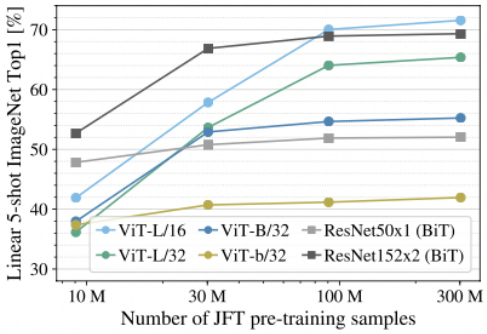


Figure 4: Linear few-shot evaluation on ImageNet versus pre-training size. ResNets perform better with smaller pre-training datasets but plateau sooner than ViT, which performs better with larger pre-training. ViT-b is ViT-B with all hidden dimensions halved.

▼ Scaling Study

- JFT를 기반으로 pre-training cost 대비 transfer 성능을 검증하여 모델들의 scaling study를 진행함
- ViT가 성능과 cost의 trade-off에서 ResNet(BiT)보다 우세한 것을 검증함
- Cost가 증가할수록 Hybrid와 ViT의 성능과 cost의 trade-off 차이가 감소함

> 💡 Pre-training cost: TPUv3 accelerator에서 모델의 inference 속도 관련 지표

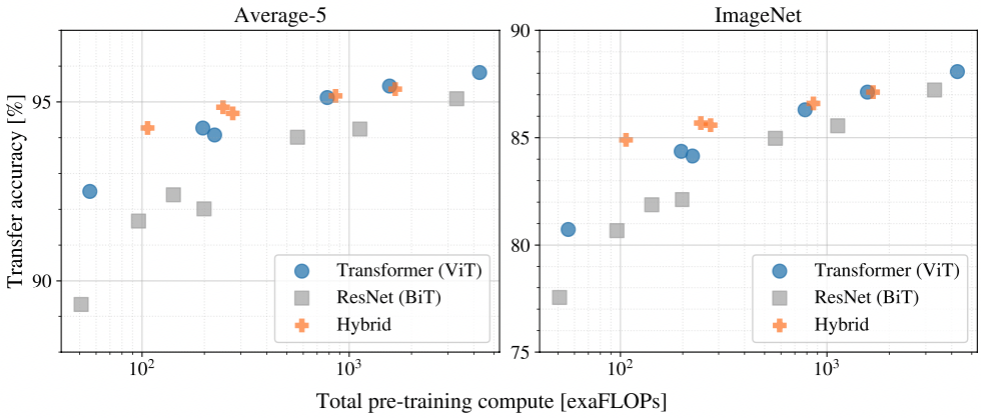


Figure 5: Performance versus pre-training compute for different architectures: Vision Transformers, ResNets, and hybrids. Vision Transformers generally outperform ResNets with the same computational budget. Hybrids improve upon pure Transformers for smaller model sizes, but the gap vanishes for larger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

▼ Inspecting Vision Transformer

- 본 실험에서는 ViT가 어떻게 이미지를 처리하는지 이해하기 위한 실험을 진행함
- (왼쪽) flatten 패치를 패치 임베딩으로 변환하는 linear projection의 principal components를 분석함
- (중간) 패치간 position임베딩의 유사도를 통해 가까운 위치에 있는 패치들의 position임베딩이 유사한지 확인함
- (오른쪽) ViT의 layer별 평균 attention distance를 확인한 결과, 초반 layer에서도 attention을 통해 이미지 전체의 정보를 통합하여 사용함을 알 수 있음

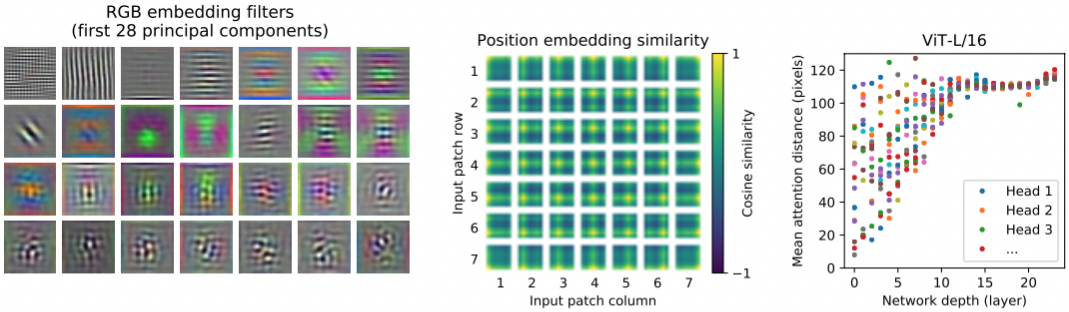


Figure 7: **Left:** Filters of the initial linear embedding of RGB values of ViT-L/32. **Center:** Similarity of position embeddings of ViT-L/32. Tiles show the cosine similarity between the position embedding of the patch with the indicated row and column and the position embeddings of all other patches. **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.7 for details.

[Paper Review] ViT: An Image is Worth 16x16 Words:Transformers for Image Recognition at Scale

ABSTRACT

기존까지 Transformer Architectures는 NLP에서는 실질적인 표준이 되었지만. 컴퓨터 비전 영역에서는 여전히 제한적으로 사용되었다.

본 논문에서는 CNN구조를 사용하지 않고 오로지 Transformer로만 사용하며, sequences of image patches에 직접적으로 적용하여 이미지 분류 작업에서 매우 잘 수행 되었다고 한다.

그리고 CNN의 SOTA와 비교해서 우수한 결과를 얻으면서도 훈련하는데 필요한 리소스가 상당히 적었다.

1. INTRODUCTION

NLP에서 사용되는 **standard Transformer**를 이미지에 그대로 적용하여 Vision Transformer(ViT)를 제안함

이미지를 패치로 분할한 후, NLP의 단어로 취급하여 각 패치의 linear embedding을 순서대로 Transformer의 input으로 넣어 이미지를 분류

훈련 시 비슷한 크기의 ResNet보다 낮은 정확도를 도출하는것을 통해 ViT가 CNN보다 inductive bias가 낮은 것을 알 수 있음

large scale데이터를 pre-training 하는 것으로 inductive bias로 인한 성능 저하를 해소할 수 있음

NLP에서 Transformers의 computational efficiency and scalability 덕분에 100B 가 넘는 parameters 크기의 모델까지 train할 수 있었음.

CNN을 완전히 대채하려는 모델을 만들었는데(Ramachandran et al., 2019; Wang et al., 2020a) 이 모델은 이론적으로 효율적이지만 specialized attention patterns 때문에 현대의 하드웨어 가속기에서는 효과적으로 확장되지는 않았음.

따라서 대규모 이미지 인식에서 ResNet과 같은 아키텍처가 여전히 최첨단이다

NLP에서 Transformer scaling의 성공에서 영감을 받아 가능한한 적은 수정으로 standard Transformer를 Image에 직접적으로 적용하는 실험을 하였다.

이를 위해, 이미지를 패치로 분할하고 이러한 패치의 sequence of linear embeddings를 Transformer의 Input으로 제공했다

이러한 이미지 패치는 NLP Application의 token(words)와 동일한 방식으로 다룸.

그리고 이미지 분류 모델을 supervised learning으로 학습 시켰음.

강력한 정규화 없이 Image Net과 같은 중간 규모의 데이터 세트를 Train할때, 이러한 모델은 유사한 크기의 ResNet보다 몇 퍼센트 낮은 정확도를 제공 하였음

하지만 모델이 더 큰 데이터 셋에서 훈련되면 달라짐.

ViT는 충분한 스케일로 pre-trained되고 datapoints(?)가 적은 작업으로 전송 될때 우수한 결과를 얻었음. ImageNet-21k dataset 나 in-house JFT-300M dataset로 pre-trained할때 multiple image recognition에서 Sota를 달성함

best model로 ImageNet에서 88.55%, ImageNet-ReaL에서 90.72%, CIFAR-100에서 94.55%, **Visual Task Adaptation Benchmark (VTAB)의 19가지 tasks에서 77.63%를 달성함**

2. Related Work

Transformers는 기계 번역을 위해 제안되었으며 이후 많은 NLP tasks에서 최첨단 방법이 되었음. Large Transformer-based models은 흔히 large corpora(말뭉치)에서 pre-trained되고, 주어진 과제에 맞게 fine-tuned됨.

Bert - denoising self-supervised pre-training task인 반면

GPT - language modeling as its pre-training task

Naive하게 self-attention을 이미지에게 적용하려면 각 픽셀이 모든 픽셀을 attends 해야 하기 때문에. 픽셀수가 quadratic cost인 경우, 이는 현실적인 규모의 input size가 아님

따라서 이미지 처리에서 Transformer를 적용시키기 위해, 과거에 몇몇의 approximations(근사치)로 시도해 보았음

Parmar et al. (2018)는 globally가 아닌 local neighborhood에만 각 query pixel에 대해 self-attention을 적용했음 이러한 local multi-head dot-product self attention blocks은 컴볼루션을 완전히 대체 할 수 있었음

다른 작업 라인에서, Sparse Transformers (Child et al., 2019)는 이미지에 적용할 수 있도록 global self-attention를 할 수 있도록 scalable approximations(확장 가능한 근사치)를 사용했음.

scale attention를 대체 하는 방법은 다양한 크기의 blocks를 적용 시키는 것임. 극단적인 경우 개별 축을 따라서만 적용함(along individual axes)

이렇게 많은 specialized attention 아키텍쳐는 컴퓨터 비전 task에서 좋은 결과를 보여주지만 hardware accelerators에서 효율적으로 구현되기 위해서는 복잡한 엔지니어링이 필요했음.

Cordonnier et al. (2020) Vit와 가장 연관이 높은데, input image로 부터 2 × 2 크기의 패치를 추출하고 위에 full self-attention를 적용하는 것인데, ViT와 매우 유사하지만, ViT는 더 나아가서 large scale pre-training이 평범한 Transformers를 SOTA CNN에 뒤지지 않거나 (더 나은)것을 입증함.

그리고 Cordonnier et al. (2020)는 2 × 2 픽셀의 작은 패치 크기를 사용하여 모델이 저해상도 이미지에만 적용할 수 있었는데, ViT는 중간 해상도 이미지도 처리 할 수 있음.

다른쪽에선 image classification를 위한 augmenting feature maps 하거나 self-attention을 사용하여 CNN의 출력을 추가로 처리 함으로서 CNN을 self-attention 형태와 결합 하는데 많은 관심이 있었음. 예를 들어(object detection, video processing, image classification, unsupervised object discovery, or unified text-vision tasks).

또 다른 최신 관련 모델은 이미지 해상도와 color space를 줄인 후 이미지 픽셀에 Transformers를 적용한 모델(iGPT) (Chen et al., 2020a)이 있었음. 이 모델은 generative model(생성모델)로 비지도 방식으로 훈련되며, 그로부터 얻어진representation는 classification performance를 위해 fine-tuned되거나 probed linearly(선형 탐색?)하여 ImageNet에서 72%의 최대 정확도를 달성할 수 있었음.

본 연구는 “표준 ImageNet 데이터셋보다 더 대규모의 이미지 recognition 데이터셋에 대한 탐구"에 관한 논문들에 결을 같이한다고 볼 수 있다. 추가 데이터 소스를 사용하면 표준 벤치마크에서 SOTA를 달성할 수 있다.

또한, Sun et al. (2017)과 같이 데이터 세트 크기에 따라 CNN 성능이 어떻게 변화하는지 연구하고, Djolonga et al. (2020). ImageNet-21k 과 JFT-300M 같은 대규모 데이터 세트에서 CNN transfer learning에 대한 empirical exploration(경험적 탐색)하는 연구도 있었다.

본 연구는 이러한 두 개의 데이터 세트에도 중점을 두지만, 이전 작업에서 사용된 ResNet 기반 모델 대신 트랜스포머를 훈련시킨다.

3. METHOD

트랜스포머 인코더는(Attention is all you need)과 최대한 유사하게 하려고 하였다. 이렇게 의도적으로 simple한 setup의 이점은 쉽게 확장이 가능한 NLP transformer 구조와 그것의 효율적인 구현을 거의 바로 사용할 수 있다는 것이다

3.1 VISION TRANSFORMER(ViT)

모델의 개요는 그림 1과 같음. 표준 Transformer은 input으로 1D sequence of token embeddings를 받는데 ViT에서는 2D 이미지를 처리하기 위해 3차원 이미지($x \in \mathbb{R}^{H \times W \times C}$)를 2차원 패치 ($x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$)로 Reshape함.

(H, W)원본 이미지의 해상도, C 이미지 채널 수, (P, P) 패치들의 해상도(논문에서는 16, 16 이겠지). N은 패치의 수를 의미 하기 때문에 $N = NH/P^2$임. 이것은 Transformer Encoder에 input 되는 유효 시퀀스 길이로 볼 수 있을 것임. 트랜스포머의 모든 레이어에 사이즈가 D인 latent vector가 사용하기 때문에, 저자들은 2차원인 patch($x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$)들을 다시 1차원으로 flatten($x \in \mathbb{R}^{N \cdot P^2 \cdot C}$)시킨후 trainable linear projection을 거쳐 D차원으로 매핑시켰음.

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

논문의 저자들은 이 projection의 결과 벡터를 patch embedding이라고 부르기로 하였음

BERT의 [CLS] 토큰과 같이 저자는 학습 가능한 class 토큰 임베딩 벡터를 Transformer encoder의 ouput 상태(z_L^0)가 이미지 표현 y

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

로 작용하는 embedded patches 시퀀스 앞에 추가하였음 ($z_0^0 = x_{class}$). (**임베딩된 패치들의 맨 앞에 하나의 학습가능한 class 토큰 임베딩 벡터를 추가)**이 임베딩 벡터를 classfication head라고 함. classfication head는 pre-training시 하나의 hidden layer를 가진 MLP로 구현하고, fine-tuning때는 single linear layer으로 나타냄.

patch embeddings 위치 정보를 유지하기 위하여 **학습가능한 1D position embeddings을 patch embeddings에 더해준다**. 저자들은 advanced한 2D-aware position embeddings이 오히려 1D position embeddings보다 유의미한 성능향상을 가져다 주기 않았기 때문이라고 한다. 이렇게 최종 구성된 임베딩 벡터 시퀀스가 encoder의 input으로 들어간다.

트랜스포머 인코더는 multi-headed self-attention(MSA) layer들과 MLP 블록들이 교차되어 구성된다.

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

Layer normalization이 모든 block의 전에 적용되며, residual connection이 모든 블록 이후에 붙는다. MLP는 GELU(Gaussian Error Linear Unit)를 activation으로 사용하는 2개의 layer를 포함하여 구성됨.

Inductive bias(귀납편향)

저자는 Vision Transformer가 CNN보다 이미지별 Inductive bias가 훨씬 적다는 것에 주목하였다. CNN에서 locality, two-dimensional neighborhood structure, 그리고 translation equivariance들은 전체 모델에서 각 layer로 baked(?)된다. (위 요소들이 CNN모델의 모든 layer에서는 내재되어 있다는 듯)

ViT에서 MLP계층만 local하고 translationally equivariant함. 반면에 self-attention layer은 global함.

two-dimensional neighborhood structure는 매우 적게 사용됨.

- 모델의 시작 부분에서 이미지를 패치로 잘라내고, fine-tuning시 다른 해상도에 대해 position embeddings을 조정할 때(Vit는 성능 향상을 위해 fine-tuning시 pre-trainig때의 이미지 보다 더 고해상도 이미지를 사용 함)를 제외하고 모델을 초기화 할 때 position embedding 이미지 패치의 2D position에 대한 정보를 전달하지 않고, 패치 간의 모든 spatial relations(공간 관계)를 처음 부터 학습 해야 한다.

Hybrid Architecture

raw image patches의 대안으로, 입력 시퀀스는 CNN의 feature maps에서 형성될 수 있음 (LeCun et al., 1989). (가공 되지 않은 이미지 패치 -픽셀로 이루어진- 에 대안으로 feature maps을 입력으로 사용 될 수 있다는 의미인 듯) Hybrid model에서 patch embedding

projection - E-는 CNN feature map에서 추출한 Patches가 적용 됨. 특별한 경우 pathces는 spatial size 1×1 를 가질 수 있는데, 이것은 입력 시퀀스가 단순히 feature maps의 spatial dimensions을 flattening하고 Transformer dimension에 projecting함으로서 얻어지는 것을 의미함. classification input embedding과 position embeddings는 상술한 바처럼 추가된다.

3.2 FINE-TUNING AND HIGHER RESOLUTION

일반적으로, 대규모 데이터 세트에서 ViT를 Pre-train하고 downstream tasks에 맞게 fine-tune함. 이를 위해, 저자는 pre-trained된 prediction head를 제거하고 0으로 초기화 된 $D \times K$ feedforward layer를 부착함. (K는 downstream classes의 갯수)

fine-tuning시 pre-training 보다 높은 해상도로 fine-tune하는게 종종 효과적임. 그렇다면 더 높은 해상도로 이미지를 제공할때, patch size가 pre-traning과 fine-tuning이 같다면 effective sequence가 더 길어질 것임. $(N = NH/P^2))$이니깐! ViT는 임의의 시퀀스 길이(하이퍼파라미터겠지?)를 조절할 수 있겠지만, pre_trained의 포지션 임베딩의 더이상 의미를 잃을 것임(fine-tuning과 pre-trained의 입력 시퀀스 길이가 다르니깐!) 그러므로 이를 극복하고자 저자는 original image에서의 위치에 따라 position embeddings을 2D interpolation을 수행함.

위와 같은 해상도 조정 및 patch extraction이 image의 2차원 구조에 대한 inductive bias가 ViT에 수동적으로 주입되는 유일한 과정임(?).

4. Experiments

연구진들은 ResNet, ViT, 그리고 하이브리드 모델의 representation learning capabilities를 평가했다. 각 모델의 data requirements를 이해하기 위해, 다양한 사이즈의 데이터셋으로 사전훈련을 진행하였고 많은 벤치마크 테스크에 대해 평가하였다. 모델의 사전훈련 계산 비용 측면에서 ViT가 다른 모델들보다 더 낮은 pre-training 비용으로 SotA를 달성하여 이 점이 있었다. 끝으로 self-supervised를 이용한 작은 실험을 수행하여 self-supervised된 ViT가 유망한 가능성을 가지는것을 보였다.

4.1 SETUP

Datasets

모델의 scalability를 탐색하기 위해, ImageNet, ImageNet-21k, JFT 등 다양한 스케일의 데이터셋에 대해 각각 모든 후보모델들을 pre-train하고 다양한 벤치마크 task들에 대해 전이하여 모델 성능을 측정하였다. 벤치마크 데이터셋으로는 ImageNet(사전훈련때 사용한 데

이터들은 제외하고 남은 holdout set으로 평가), ImageNet ReaL, CIFAR-10/100, Oxford-IIIT Pets, Oxford Flowers-102 등을 이용하였다.

Model Variants

저자는 ViT configurations은 BERT에 사용된 configurations을 base로 삼았으며 'Base'와 'Large'모델은 BERT와 완전히 동일하다. 여기에 더 큰 규모의 'Huge'모델을 추가하였다. notation : ViT-L/16 → 16*16 patch size를 사용하는 Large variant 모델. (패치 사이즈와 transformer input sequence의 길이는 반비례하므로 패치 크기가 작을수록 계산 코스트가 증가)

baseline CNN으로 ResNet을 사용하였는데, 오리지날 ResNet의 Batch Normalization을 Group Normalization으로 대체하였고 standardized convolution을 사용하였다. 이러한 작은 변형이 전이학습의 성능을 더 향상시키며, 이 modified model을 ResNet(BiT)로 표기한다. 하이브리드 모델에서는 CNN의 중간 feature map들을 ViT에 1*1 패치로 쪼개어 넣어주었다.

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

사전훈련시에는 모든 모델을 Adam 옵티마이저로 훈련시켰다.. 하이퍼파라미터 세팅은 $\beta_1 = 0.9, \beta_2 = 0.9999,$ batch_size=4096로 두었고, weight decay를 적용하였는데 이것이 모든 모델에 대해 transfer시 유용한 도움을 주는것을 발견하였다. 학습률 스케쥴로는 linear learning rate warmup and decay를 사용하였다. 파인튜닝시에는 SGD with momentum 옵티마이저를 사용, batch_size = 512으로 두고 훈련하였다.

Metrics

저자들은 downstream 데이터셋에 대한 결과를 few-shot accuracy 혹은 fine-tuning accuracy로 보고하였다. Fine-tuning acc는 각 모델의 각 데이터셋에 대한 퍼포먼스를 반영한다. Few-shot accuracy는 훈련이미지들의 subset의representation을 $\{-1, 1\}^K$ 의 타겟벡터로 매핑하는 규제된 linear regression 문제의 해를 구하는 과정에서 얻어진다. 주로 fine-tuning performance에 집중하였으나 때때로 fine-tuning이 너무 고비용인 경우 가벼운 평가가 가능한 few-shot accuracy를 사용하였다.

4.2 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

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).

저자들은 먼저 ViT variants 모델들 중 가장 큰 size의 ViT-H/14와 ViT-L/16을 문헌기준으로 SotA에 자리매김하고있는 CNN들과 비교해보았다. 첫 비교대상은 대규모의 ResNet으로 지도방식의 전이학습을 수행시킨 Big Transfer(BiT)모델이고, 두 번째 대상은 큰 EfficientNet으로 준지도학습을이용해 학습시킨 Noisy Student이다. 본 연구가 진행되던 시점 기준 ImageNet에 대해서는 Noisy Student가 SotA이고, 다른 데이터셋들에 대해서는 BiT-L이 SotA이다. 모든 실험모델들은 TPUv3에서 훈련되었고 각각을 사전훈련하는데 소요된 일 수와 가동 코어 수를 곱한값을 TPUv3-core-days라고 표현하였다.

Table2를 보면 동일한 데이터셋에 대해 사전학습되었을 때, 모든 downstream task에서 ViT-L/16의 성능이 BiT-L을 능가하면서 동시에 사전훈련에 드는 계산상의 cost는 확연하게 낮음을 확인할 수 있다. 더 큰 모델인 ViT-H/14은 모든 실험 데이터셋에 대해 더욱 향상된 성능을 보여준다. 특히 더욱 도전적인 데이터셋인 ImageNet, CIFAR-100, VTAB 세트 등에서 두드러진다. 흥미롭게도 이 Huge한 모델도 이전의 SotA들과 비교했을 때 더 낮은 pre-train 비용이 요구된다. 그러나 연구진들은 사전학습의 효율성이 데이터 구조의 선택뿐 아니라 training schedule이나 optimizer, weight decay등의 다른 파라미터들에도 영향을 받는다는 것에도 주목하였다. performance와 compute에 대하여 다른 구조들의 모델들에 대해 제어된 실험을 실시한 결과는 Section 4.4에서 확인할 수 있다. 끝으로 ViT-L/16을 JFT보다 작은규모의 데이터셋인 ImageNet-21k에 pre-train 하였을 때에도 대부분에 데이터 셋에 대해 상당히 좋은 성능을 보이는 것을 확인하였다. (비교적 적은 pre-train resource에도 불구하고 : 표준 클라우드의 8코어 TPUv3로 대략 30일 정도면 훈련 가능!)

4.3 PRE-TRAINING DATA REQUIREMENTS

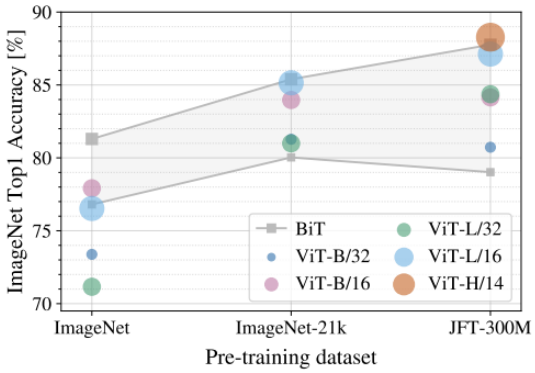


Figure 3: Transfer to ImageNet. While large ViT models perform worse than BiT ResNets (shaded area) when pre-trained on small datasets, they shine when pre-trained on larger datasets. Similarly, larger ViT variants overtake smaller ones as the dataset grows.

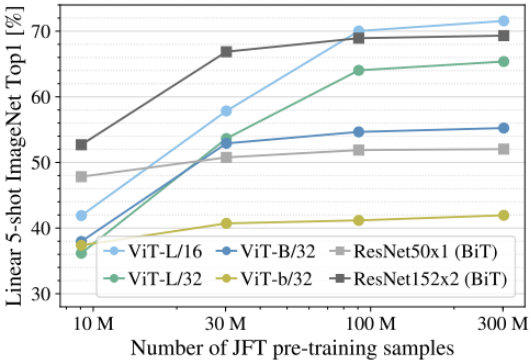


Figure 4: Linear few-shot evaluation on ImageNet versus pre-training size. ResNets perform better with smaller pre-training datasets but plateau sooner than ViT, which performs better with larger pre-training. ViT-b is ViT-B with all hidden dimensions halved.

Vision Transformer는 대규모의 JFT-300M 데이터셋에 사전학습될 때 기존의 SotA ResNet들보다 더 적은 inductive biases들을 가지고도 더 잘 작동한다. 그런데 이때 데이터셋의 '사이즈'가 얼마나 중요할까? 다음과 같은 두 가지 실험을 시도해봤다고 한다.

먼처 첫 번째로 ViT를 데이터셋의 사이즈를 증가시켜보면서 (ImageNet → ImageNet-21k → JFT-300M) 사전학습 시켰다. 작은 데이터셋에도 가능한 최고의 성능을 얻게하기 위해 규제파라미터(weight decay, dropout & label smoothing)를 optimize 하였다고 한다. Figure3은 각 모델들을 각 데이터셋에 pre-train시킨 뒤 ImageNet에 fine-tuning한 성능측정 결과이다. 가장 작은사이즈의 표준 ImageNet에 사전학습 시켰을 때는 빡센 규제에도 불구하고 ViT-Large가 ViT-Base보다 성능이 안나온다. 조금 더 큰 규모의 ImageNet-21k에 사전훈련 시킨 경우에도 비슷하다. 오직 JFT-300M(가장 대규모)으로 pre-train 했을 때 큰 모델의 완전한 이점을 확인할 수 있었다. Figure3에서 색칠된 회색 영역은 BiT-50, 152의 성능 구간인데, 작은 데이터셋으로 사전훈련했을 때에는 BiT가 앞서지만 사전훈련 데이터셋 규모가 커질수록 ViT가 BiT를 능가해간다.

두 번째 실험으로써, 저자들은 full JFT-300M 데이터셋 뿐만아니라 그것의 랜덤한 부분집합(9M, 30M, 90M)에 대해서도 훈련시켜보았다. 작은 부분집합에 훈련시킬 때에도 기본 JFT에 대해 훈련시 사용한 하이퍼파라미터들을 그대로 사용하였다. 이를 통해 규제의 효과가 아닌 모델의 본질적인 성질들을 평가하였다. 이 실험에서는 계산자원을 절약하기 위해 일반적인 full fine-tuning accuracy를 보는 대신 few-shot linear accuracy를 모형평가의 측도로써 사용하였다. Figure4가 그 결과를 보여주는데, 상대적으로 작은 데이터셋에 대해서 사전훈련시에는 역시 BiT가 ViT를 앞었다. ViT는 이런 소규모 데이터셋에 대해 BiT보다 더 많이 overfit되었다. 예를들면 ResNet 50보다 ViT-B/32가 살짝 더 빨랐지만, 9M보다 작은 데이터셋에 대해서 성능이 더 낮았으며, 90M보다 큰 데이터셋에 대해서는 성능이 더 나았다. 이 양상은 ResNet 152 vs ViT-L/16의 비교에서도 동일하게 나타났다. 이 결

과는 'convolutional inductive bias'가 상대적으로 작은 데이터셋에 대해서 useful하지만, 큰 데이터셋에 대해서는 단지 적절한 패턴을 학습하는것 만으로 충분하거나 심지어 더 이롭다는 직관을 공고하게 만들어주었다.

비전트랜스포머의 few-shot properties의 추가적인 분석은 하나의 흥미로운 향후 연구방향이 될 것 같다.

4.4 Scaling Study

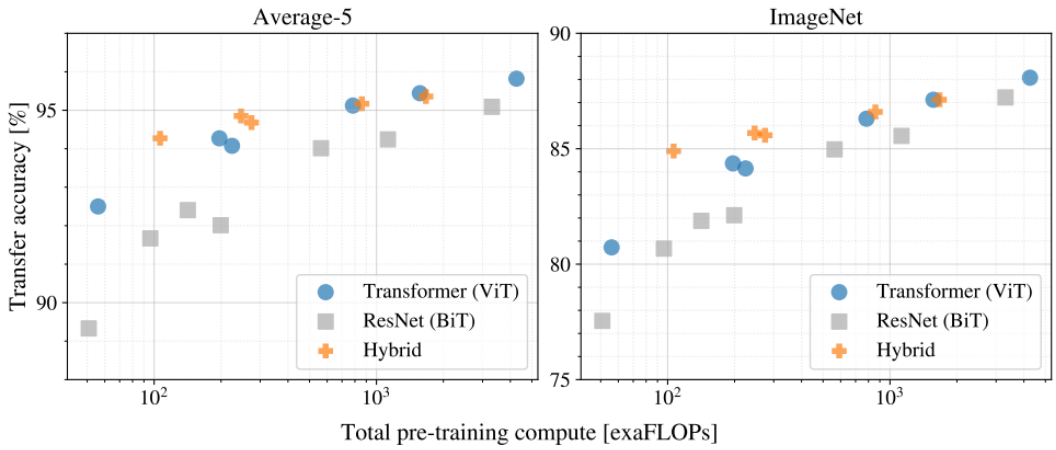


Figure 5: Performance versus pre-training compute for different architectures: Vision Transformers, ResNets, and hybrids. Vision Transformers generally outperform ResNets with the same computational budget. Hybrids improve upon pure Transformers for smaller model sizes, but the gap vanishes for larger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

JFT-300M데이터셋으로부터의 전이학습 성능평가를 수행하는데 각 모델의 performance vs pre-training cost를 평가하였다. ResNet 7개와 ViT 6개, 하이브리드모델 5개의 모델들을 고려하였다. Figure5를 통해 몇 가지 패턴을 확인할 수 있는데,

첫 번째로 모든 비전트랜스포머 모델이 ResNet 모델들을 성능/계산코스트 trade-off에서 압도한다는 것이다. 동일한 성능을 달성하기위해 드는 계산비용이 ViT가 2 ~ 4배는 더 적다.

두 번째로 하이브리드 모델이 비교적 작은 계산영역에서는 ViT의 성능을 앞질렀다는 것이다. 이 차이는 모델 사이즈가 커짐에 따라 vanishing되긴한다. 이 결과가 다소놀라운것이 어느사이즈의 ViT에도 convoluational local feature processing이 도움을 줄 수 있다는 것을 기대할 수 있게하기 때문이다.

마지막 세 번째로 ViT는 실험을 시도해본 range 내에서 saturated 되지 않았기에, 미래의 feature scaling efforts를 고무시킨다(?).

4.5 Inspecting Vision Transformer

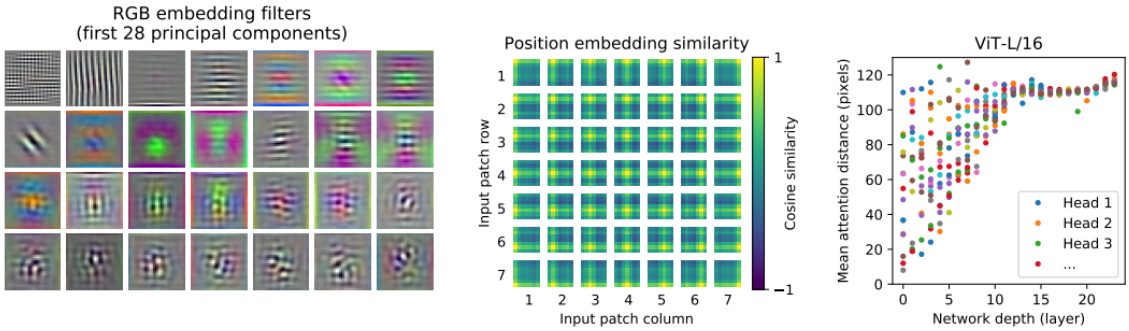


Figure 7: **Left:** Filters of the initial linear embedding of RGB values of ViT-L/32. **Center:** Similarity of position embeddings of ViT-L/32. Tiles show the cosine similarity between the position embedding of the patch with the indicated row and column and the position embeddings of all other patches. **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.7 for details.

비전트랜스포머가 이미지를 어떻게 처리하는지 이해하기 위해 내부의 representation들을 분석하였다.

ViT의 첫 번째 층은 flatten된 patch들을 저차원 공간으로 linearly project한다. Figure7의 왼쪽 사진이 해당 층에서 학습된 임베딩 필터들의 상위 주성분들을 보여주고있다. 저러한 성분들은 각 패치 내의 미세 구조의 저차원적 표현을 위한 그럴듯한 기저함수를 닮았다고 볼 수 있다.

projection이후에는 학습된 position embedding이 patch representation에 추가된다. Figure7의 가운데 그림은 모델이 포지션임베딩의 유사도 내에서 이미지 내부의 거리개념을 인코딩하는 방법을 배운다는 것을 보여준다. 즉, 가까운 패치들은 유사한 포지션 임베딩을 가지며 row-column 구조(같은 행/열에 있는 patch는 유사한 임베딩을 갖는다.) 또한 나타난다. note, 위치인코딩을 더하고 안더하고는 모델의 성능차이가 큰데 1d pos embedding과 2d pos embedding의 차이는 거의 없으며 오히려 1d가 더 낫다.

self-attention은 ViT가 이미지 전체의 정보를 통합하여 사용할 수 있게한다. (심지어 최 하위층(input 바로 다음 층)에서도) 논문저자들은 네트워크가 이 광활한 수용력을 얼마나 이용하는지 그 정도를 조사해보았다고 한다. 구체적으로는 정보가 attention weights에 의해 통합되는 image space내의 평균거리를 계산하였다(Figure7의 우측 그림). 이 "어텐션 거리"는 CNN에서의 receptive field size와 유사한 개념이라고 보면된다. 그림을 보면 우선 층이 깊어질수록 어텐션거리가 증가하며 심지어 최하위층 레이어에서도 몇몇 attention head가 이미지 대부분에 attend하고있는 것을 확인할 수 있는데, 이는 정보를 global하게 integrate할 수 있는 능력을 모델이 실제로 사용하고 있음을 보여준다. 또한, 하위층에서 다른 몇몇 어텐션 헤드들은 작은 어텐션거리를 갖는데, 이 고도로 local한 어텐션은 hybrid 모델에서는 덜하다. 이는 이 local한 어텐션이 CNN의 초기 convolutional layer들과 유사한 기능을 할 수 있음을 알려준다.

또한 저자들은 ViT가 실제로 분류에 의미적으로 관련있는 영역에 attend한다는 것을 발견하였다.(Figure6)

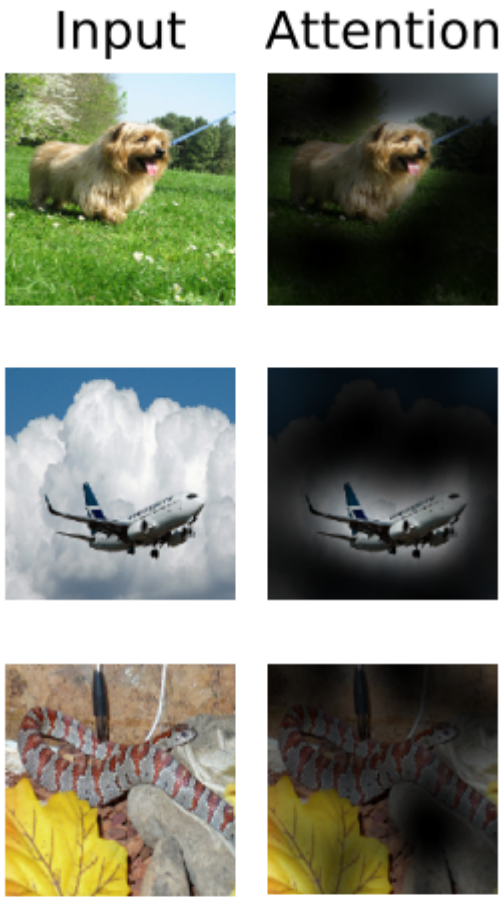


Figure 6: Representative examples of attention from the output token to the input space. See Appendix D.7 for details.

4.6 Self-Supervision

트랜스포머기반 모델들은 NLP task들에서 인상깊은 퍼포먼스를 보여주었다. 그러나 그것들의 많은 성공은 뛰어난 확장성 뿐 만아니라 대규모의 자기지도 사전훈련(self-supervised pre-training)으로부터 비롯된다. 연구진들은 BERT에서 사용된 masked language modeling task를 모방하여 masked patch prediction for self-supervision를 실험해보았고, 그 결과는 사전학습을 하지 않았을 때에 비해 유의미한 성능향상을 가져다주었

다. 그러나 여전히 지도방식의 사전훈련(supervised pre-training)에는 많이 못미치는 성능이었다. 이러한 self-supervised pre-training에 대한 탐험은 미래의 연구로 남겨두고있다.

5. Conclusion

저자들은 Transformer를 이미지 인식에 직접 적용하는 방법을 탐구했다. 비전분야에 self-attention을 사용하는 이전의 연구들과는 다르게 본 연구에서는 모델을 설계할 때, 이미지에 특화된 그 어떠한 inductive bias도 추가하지 않았다. 대신에 그들은 이미지를 patch들의 시퀀스로 보고 NLP에서의 표준 트랜스포머의 encoder 부분을 이용해 처리하였다. 이 간단하면서도 확장성 좋은 전략은 **큰 데이터 셋에 대한 사전훈련이 곁들여질 때** 굉장히 잘 작동한다. ViT는 이미지 분류 데이터셋에 대해 SotA수준의 성능을 찍으면서도 사전훈련이 비용이 상대적으로 저렴하다.

이러한 결과들이 고무적이지만 많은 도전과제들이 남아있다.

1. detection이나 segmentation 등의 다른 비전분야 테스크들에의 적용(동시기의 Carion에 의한 논문이 이부분의 가능성을 보여줌)
2. self-supervised pre-training 방법의 향상(본 논문의 실험들에서 이에 대한 개선을 보여주었으나, large-scale supervised pre-training과 비교했을 때 아직 격차가 심함)
3. 성능 향상을 위한 ViT의 추가적인 확장

[참고]

[ViT 시리즈] Vision Transformer 논문 리뷰 보단 메모.

Group Normalization

Vision Transformer (1)

데이터 전처리의 피처 스케일링(Feature Scaling)

머신러닝 용어: Example, Sample & Data Point