📖

[논문리뷰] Vision Transformer [2021]

🌟 본 논문에서는 **Vision에 Transformer Architecture를 활용**한, **Vision Transformer를 소개**합니다.

Introduction

Transformer의 기반이 되는 **Self-attention는 NLP 분야에서 계속해서 활용**되고 있습니다.
주된 접근방식은은 **large text corpus에 대해 pre-train**한 후, **소규모 task**에 **fine-tune**을 하는 것이다.

Transformer의 계산효율성과 학장성 덕분에, 100B의 파라미터가 넘는 이전에 생각지도 못한 크기를 모델이 학습하는 것이 가능해졌습니다.
뿐만 아니라, **모델과 데이터가 커져도, Saturating performance가 없습니다.**

NLP내에서 Transformer의 성공에 영감을 받아, 약간의 수정을 한 후, Transformer를 image에 적용하였습니다.

Transformer를 image에 적용하기 위해서, **image를 patch 단위**로 나누고, **이러한 patches들의 선형적인 embedding Sequence는 Transformer의 Input으로 활용**하였습니다.
(Image Patch들은 NLP의 token이라고 보면 됩니다.)

하지만, ImageNet의 중간 사이즈 Dataset을 ResNet과 비슷한 규모의 Model로 학습하였을 때,
정확도가 낮았습니다. 이러한 결과로 낙담할 수 있습니다.

> 💡 **등분산과 지역화와 같은 역할**을 하는 Inductive biases가 CNN에 내재되어있지만, Transformer에는 없습니다.
> 그리하여, **불충분한 데이터로 학습**을 할때, **일반화하는 것이 불가능**합니다.

✨ **그러나, Large scale training을 하는 경우, inductive bias를 능가하는 것을 발견**하였습니다.
충분한 규모의 데이터를 학습하고, **적은 데이터의 지점으로 전환**할 때,
Vision Transformer(ViT)는 훌륭한 결과를 달성하였습니다.

Related Work

Transformer는 machine translation분야에서 제안되었으며, 많은 NLP task에서 SOTA를 달성하였습니다.
대규모 거대한 데이터로 pre-trained한 후, 당면한 과제에 맞게 fine-tuned 되었습니다.

단순하게 이미지에 Self-Attention을 적용하는 것은 각 픽셀이 다른 픽셀에 attention 될것이 요구됩니다.
하지만, **이는 Pixel 수에 대해 Quadratic한 Complexity**를 가지며,
이로 인하여, 현실적으로 여러 Input size로 확장되는 것이 불가능합니다.

여러가지 **많은 실험들[local self-attention, sparse attention, block of varying size]**을 해왔지만,
하드웨어에서 실행되기 위해서는, 복잡한 엔지니어링이 요구됩니다.

연구에서는, **ImageNet보다 큰 Dataset을 활용하여, SOTA들 달성**하였습니다.
더 나아가, **dataset size에 따라 어떻게 달라질 수 있는지에 대한 연구와 대규모 데이터셋에 대한 CNN의 전이학습을 연구**하였습니다.

Method

Vaswani의 Original Transformer에서 영감을 받아,
쉽게 확장 가능한 NLP Transformer구조와 유사하게 모델을 생성하였습니다.

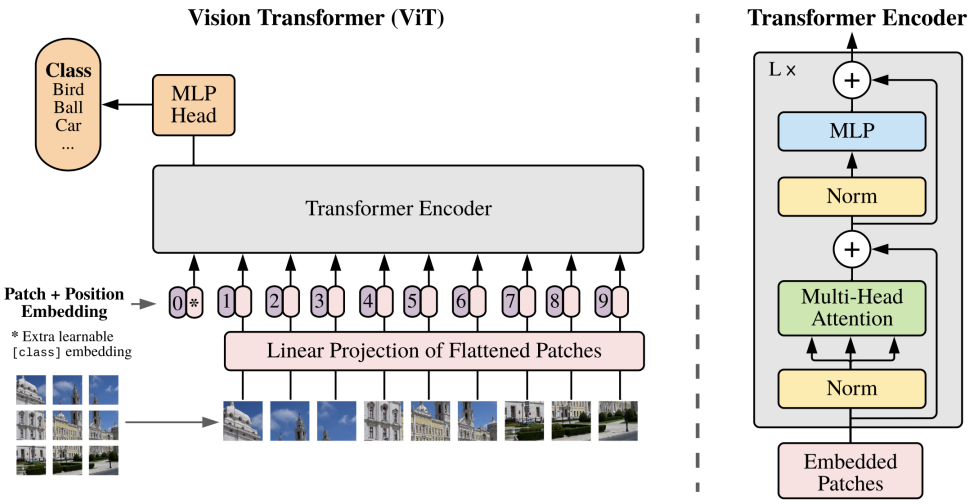


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification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

Vision Transformer [ViT]

모델의 구조는 위의 그림과 같습니다.
일반적인 **Transformer는 1차원 token Embedding Sequence를 Input으로 받습니다.**

2차원 Image를 다루기 위해,
3차원 Image인 $x \in R^{H*W*C}$를 2차원 Patch인 $x_p \in R^{N*(P^2*C)}$로 flatten 하였다.
(H, W)는 원본의 해상도이며, C는 Image의 채널 수, (P, P)는 패치들의 해상도이다.

$N = HW/P^2$는 Patch의 수를 Transformer Encoder의 Input Sequence의 길이로 볼 수 있습니다.
Transformer는 내부의 모든 layer들에 흐르는 latent vector의 size가 D로 통일 되어있는데,
2차원 patch들을 다시 1차원으로 flatten하고, 이를 trainable linear Projection을 거쳐서 D차원으로 매핑하였다.

Projection의 결과를 Patch Embedding이라고 합니다.

BERT의 token과 유사하게, 학습이 가능한 토큰 임베딩을 임베딩된 패치들의 Sequence에 추가하였다.

Classification head는 **pre-trained**할 경우, **하나의 은닉층을 가진 MLP로 구현**되었고,
fine-tunning시에는 **single linear layer로 구현**되었다.

위치정보를 유지하기 위해서, **Possition Embedding에 Patch Embedding이 추가**되었습니다.
2D-position Embedding이 1D-position Embedding과 비교했을 때, **2D의 경우, 성능향상을 가져오지 않아, 1D-Position Embedding을 Encoder Vector로 사용**하였다.

Transformer Encoder는 Multuheaded self-attention layer와 MLP blocks이 교차되어 구성되었다.
LayerNorm(LN)은 모든 block 이전에 적용되었으며, 모든 block 이후에 residual connection이 적용되었습니다.
MLP는 GELU(Gaussian Error Linear Unit)를 activation으로 사용하는 2개의 layer를 포함하고 있습니다.

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

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

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

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

Inductive Bias

Vision Trnasformer는 CNN에 비해, Image-specific inductive Bias를 덜 가지고 있습니다.
Inductive Bias란, 많은 데이터에 대해 귀납적으로 문제를 풀고 싶어하고,
그 문제를 더 잘풀기 위해 설계한 모델 / 목적함수의 추가적인 가정을 말합니다.

ViT는 CNN에 비해 이미지에 특화된 1) 국소적인 2차원 부분 구조, 2) translation equivariance의 inductive bias가 부족합니다.

2차원 이웃을 고려하는 부분은 처음 입력 이미지를 패치로 구분할 때와 실험애서 살펴볼 fine-tunning 시에 사용되고, **삽입한 position embedding은 초기화시 patch별 위치에 대한 아무 정보를 제공하지 않고 처음부터 학습**되어야만 합니다.

Hybrid Architecture

raw image patches를 대신하여,
CNN의 feature map으로부터, Input Sequence를 생성하는 것이 가능합니다.

Hybrid model에서는, Patch Embedding Projection E는 CNN의 feature map으로부터 추출된 patch들에 적용됩니다. **특별한 경우, patch는 spatial size 1x1를 가질 수 있는데, 이는 Input Sequence가 단지, feature map의 공간차원들을 flattening 한 뒤, Transformer의 차원으로 projecting됩니다.**

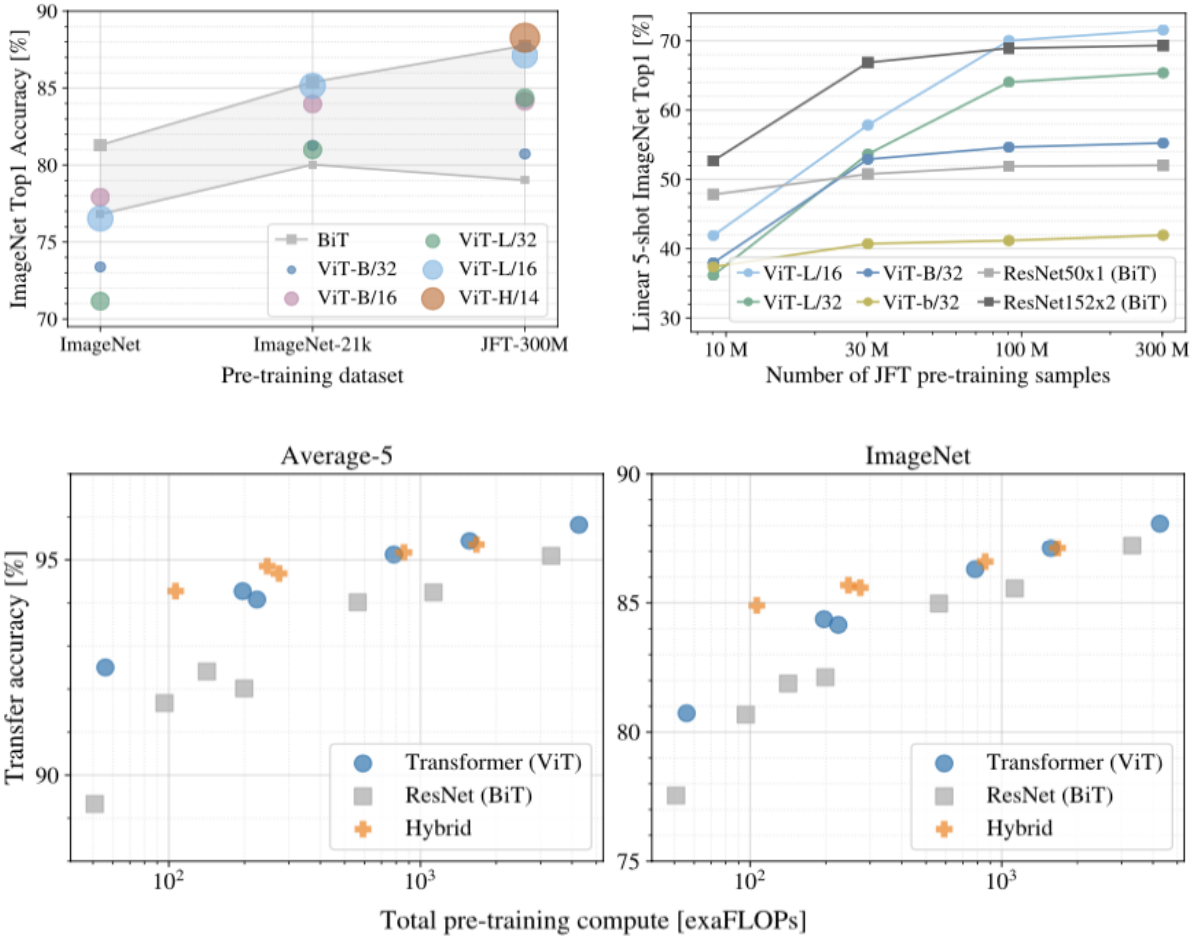
Experiments

ResNet, Vision Transformer(ViT), Hybrid의 representation learning capabilities를 평가하였다.
각 모델의 data requirements를 이해하기 위해서, 다양한 size의 Dataset으로 학습을 진행하였고, 평가하였다.

Computational cost of pre-training을 고려하였을 때, ViT는 다른 모델들보다 더 낮은 pre-training 비용으로 SOTA를 달성하였다.
또한, **self-supervision을 사용하여, 작은 실험을 하였을 때, ViT는 유망한 전망**을 보여주었습니다.

	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).



Conclusion

연구진들은 Image Recognition에서 Transformer의 적용을 연구하였다. **Image를 Patch의 Sequence로 해석하고, NLP내에서 표준적으로 사용되는 Transformer encoder를 사용**하였다. Transformer는 여러 규모로 확장 가능하며, 거대한 규모의 데이터셋에서 놀라운 성능을 보였습니다. 그리고 Vision

Transformer는 Image Classification의 여러 Dataset에서 SOTA를 달성하였습니다.
ViT를 Scaling한다면, 개선의 여지가 있음을 연구에서 보여주었다.

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

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace

https://arxiv.org/abs/2010.11929

The Illustrated Transformer

저번 글에서 다뤘던 attention seq2seq 모델에 이어, attention 을 활용한 또 다른 모델인 Transformer 모델에 대해 얘기해보려 합니다. 2017 NIPS에서 Google이 소개했던 Transformer는 NLP 학계에서 정말 큰 주목을 끌었는데요, 어떻게 보면 기존의 CNN 과 RNN 이 주를 이뤘던 연구들에서 벗어나 아예 새로운 모델을 제안했기 때문이지 않을까 싶습니다.

https://nlpinkorean.github.io/illustrated-transformer/

Vision Transformer (1)

Machine Learning/Time Series] - Transformer [Machine Learning/Time Series] - Transformer 구현 [Machine Learning/Time Series] - Transformer Positional Encoding Transformer 모델은 발표 이후에 자연어 처리 (NLP) 분야에서의 새로운

https://hongl.tistory.com/232

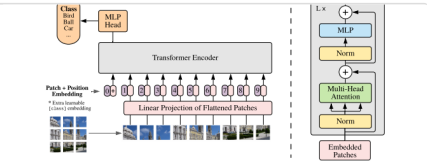


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable