AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

건국대학교
컴퓨터공학과

김동환

2024.07.01

목 차

AN IMAGE IS WORTH 16X16 WORDS:
TRANSFORMERS FOR IMAGE
RECOGNITION AT SCALE

0. Abstract

배경

- NLP에서 Transformer 아키텍쳐는 표준이 되었지만, CV에서는 CNN 기반으로 제한되어 사용되었다.

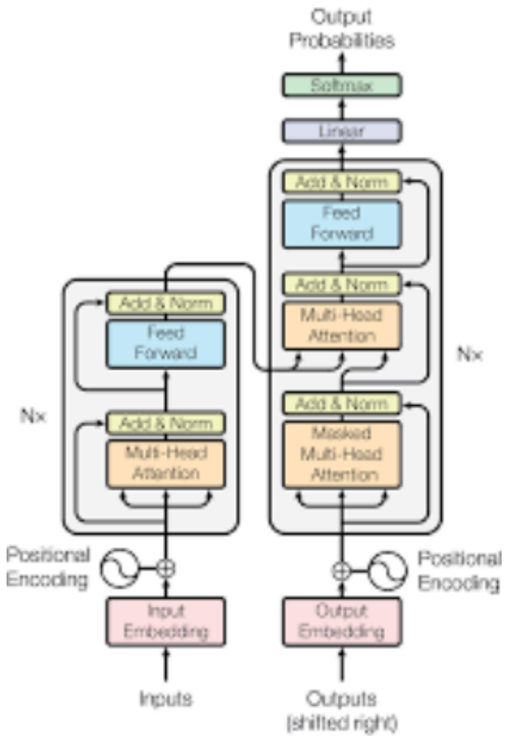
목적

- CNN 의존적인 기존의 방식에서 탈피하여 Transformer를 CV에 직접 적용하여 성능 향상을 이끌어 낼 것

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

1. Introduction

Transformer



연구 개요

NLP에서 성공적인 결과를 이끌어낸 Transformer를
CV에도 이용하기 위하여
이미지를 patch로 나누고 선형 변환하여
"이미지를 자연어와 동일하게 취급"
하는 방법을 택한다.

"AN IMAGE IS WORTH 16X16 WORDS"

AN IMAGE IS WORTH 16X16 WORDS:
TRANSFORMERS FOR IMAGE
RECOGNITION AT SCALE

2. Related Work

모든 픽셀에 Self-attention 단순 적용

O(n^2)의 시간복잡도를 가지므로 현실적이지 않음

Local Self Attention, Sparse Transformers

CNN을 대체 가능 근사 Global self-attention 적용

Block-based Attention

가변 크기의 블록 단위로 attention 적용

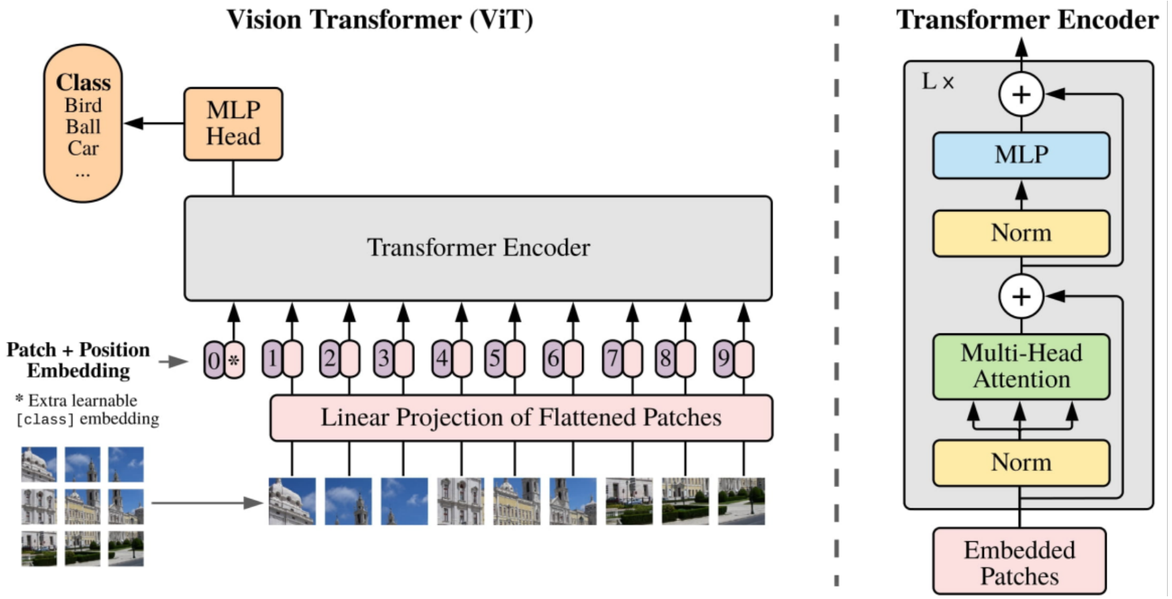
Cordonnier et al

2 x 2 patch를 사용하여 저해상도 이미지를 처리 가능, 중간 해상도 이미지 처리 불가

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

3. Method

ViT



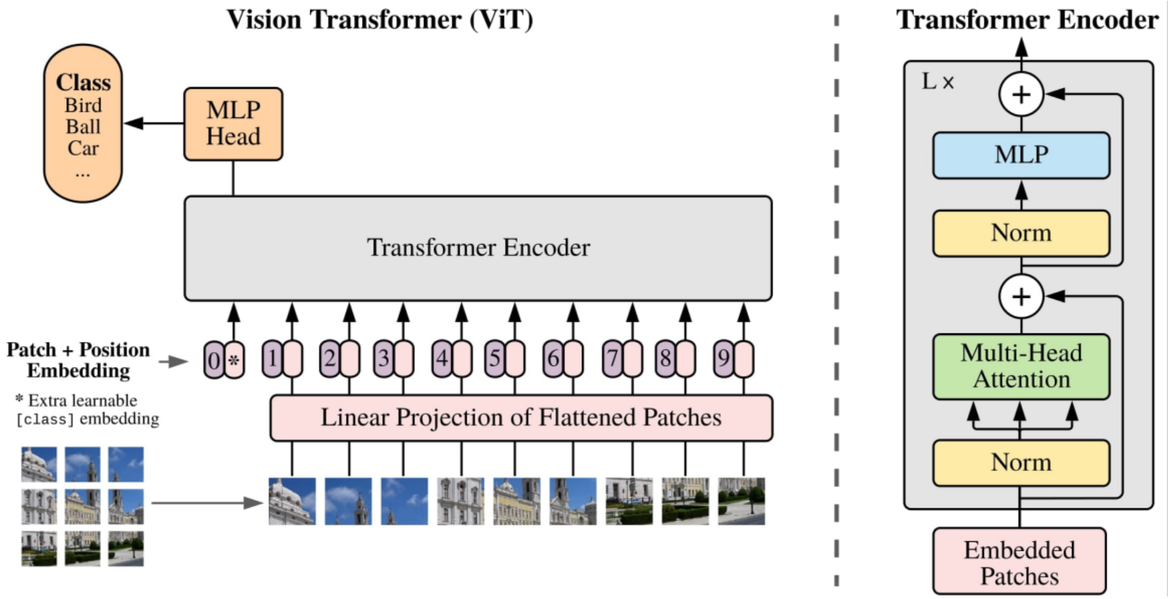
ViT의 아키텍쳐

Transformer의 확장 가능성과 구현 편의성을 위해
Transformer의 Encoder를 그대로 차용

AN IMAGE IS WORTH 16X16 WORDS:
TRANSFORMERS FOR IMAGE
RECOGNITION AT SCALE

3.1 Vision Transformer (ViT)

ViT



ViT의 Input Sequence

$$\mathbf{x} \in \mathbb{R}^{H \times W \times C} \longrightarrow \mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$$

1차원 데이터를 입력받을 수 있는 Encoder이기 때문에
x를 x_p로 가공

$$\mathbf{x}_{\text{class}}$$

Image Classification을 위해
Class token 추가

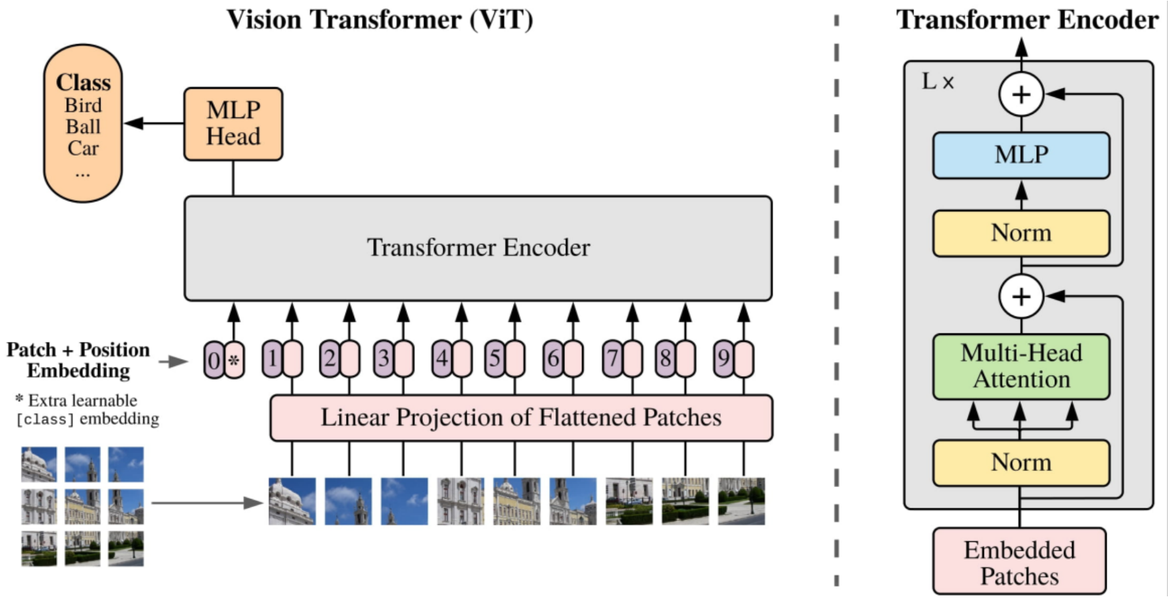
$$\mathbf{E}_{pos}$$

Patch의 위치 정보를 위한 Positional Embedding 더함

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

3.1 Vision Transformer (ViT)

ViT



ViT의 동작 방식

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

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

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

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

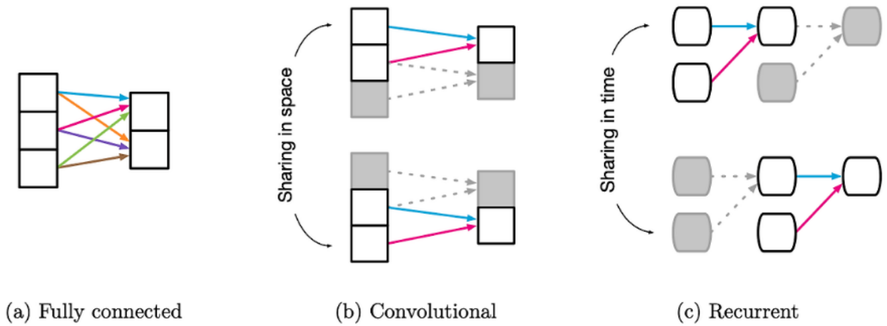
MSA와 MLP layer를 번갈가며 연산 수행
LN, residual connection은 매 단계마다 진행

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Inductive bias

CNN은 inductive bias가 높은 반면,
ViT는 낮다는 단점 존재

inductive bias



(a) Fully connected (b) Convolutional (c) Recurrent

Hybrid 아키텍쳐

Transformer에 Inductive bias를 주입하기 위하여
CNN의 feature map을 입력 값으로 사용

AN IMAGE IS WORTH 16X16 WORDS:
TRANSFORMERS FOR IMAGE
RECOGNITION AT SCALE

4. Experiments

ViT 모델과 SOTA 비교하는 실험

Model Variants

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.

BERT의 Base, Large 그대로 사용,
Huge 단위 추가

SOTA와 비교

	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

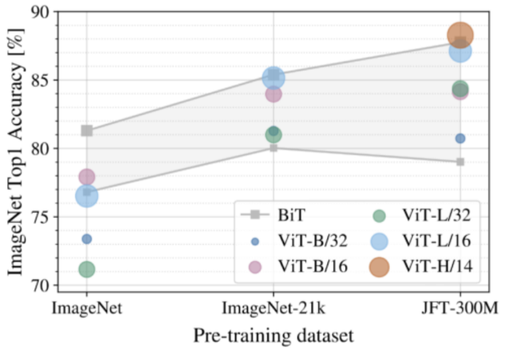
Ours(ViT)가 BiT(ResNet), Noisy Student 보다
정확도가 높을 뿐 아니라
훈련 비용이 적음

AN IMAGE IS WORTH 16X16 WORDS:
TRANSFORMERS FOR IMAGE
RECOGNITION AT SCALE

4. Experiments

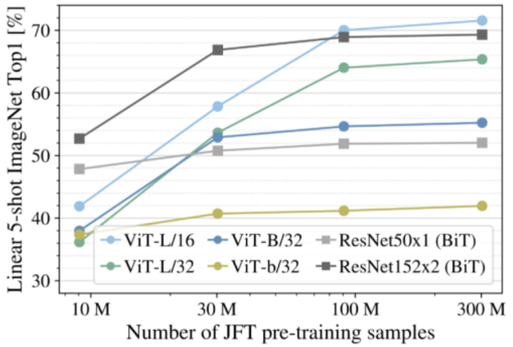
ViT는 CNN보다 inductive bias가 낮기 때문에
충분히 큰 크기의 데이터셋을 필요로 함

Dataset의 크기 조정



데이터 셋의 크기가 클수록
ViT가 CNN을 능가

Samples 크기 조정



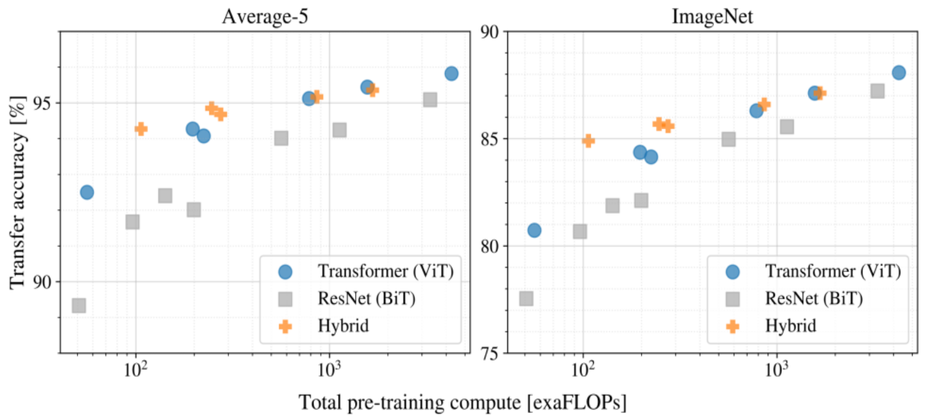
Sample이 클수록
ViT가 CNN을 능가

AN IMAGE IS WORTH 16X16 WORDS:
TRANSFORMERS FOR IMAGE
RECOGNITION AT SCALE

4. Experiments

Scaling study

pre-training cost 대비 정확도



Insight

동일 pre-training compute 일때,
ViT는 ResNet보다 우수한 성능을 보임

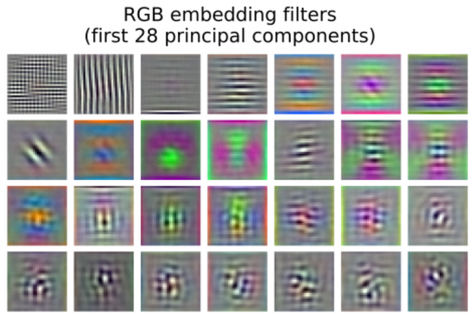
ViT는 Hybrid보다도 높은 성능을 보임

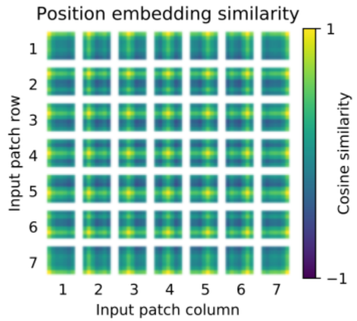
ViT는 정체되지 않은 성능을 보임

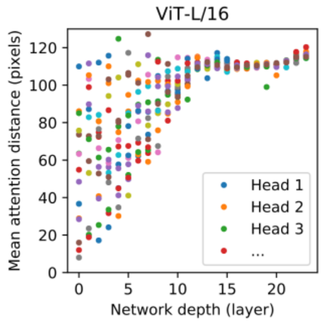
4. Experiments

Inspecting Vision Transformer

ViT 내부 표현 방식







Embedding Fiter ~= CNN Filter

Position Embedding의 지역간 유사성

ViT는 초기 layer에서도 이미지 전체를 활용하기도 함

AN IMAGE IS WORTH 16X16 WORDS:
TRANSFORMERS FOR IMAGE
RECOGNITION AT SCALE

4. Experiments

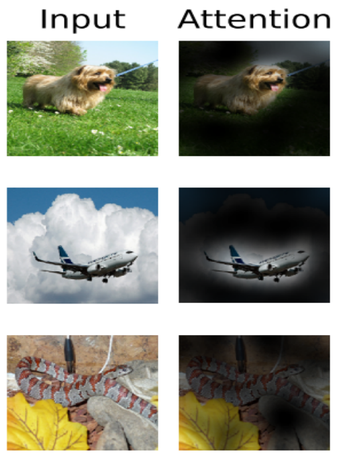


Figure 6: Representative examples of attention from the output token to the input space. See Appendix D.7 for details.

Attention을 사용하기 때문에
이미지 분류에서 어느 부분에 집중하는지
시각적으로 확인 가능

Self-supervision을 MLM 방식을 모방하여 수행한 결과
supervised learning 대비 4% 낮은 성능

5. Conclusion

결론 - 1

ViT는 기존 방식과 다르게 image-specific inductive biases를 도입하지 않음

결론 - 2

이미지를 patch로 나눈 후 Transformer Encoder에입력을 줌으로써 확장 가능성을 확보하고 SOTA 성능을 달성

결론 - 3

CNN보다 상대적으로 낮은 Inductive Bias를 데이터 양으로 극복

추후 연구과제

1. ViT를 다른 CV task에 적용
2. self-supervised learning
3. ViT를 더 크게 확장하여 성능 개선

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE