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

Dosovitskiy et al., 2020 Google Research

https://arxiv.org/abs/2010.11929

Chioh Lee
Dept. of AI, School of CSE

Abstract & Introduction

- Transformer w/ self-attention (Vaswani et al., 2017)은 NLP 분야에서 de-facto standard (dominant approach: pre-train & fine-tune)
    - Pros: computational efficiency & scalability, yet no sign of saturation
- CV 분야에선 부분적 혹은 전체를 self-attention으로 대체하려는 노력이 있었으나 복잡한 구조 때문에 hardware accelerator을 이용하기 부적절함
- 본 논문은 transformer의 원 구조를 최대한 살리고 CV 분야의 classification task에 적용, 큰 데이터 셋의 사용이 SOTA 모델들과 견줄만한 성능을 보인다는 것을 보여줌

Related works

Transformer: "Attention is all you need" (Vaswani et al., 2017)

- RNN LSTM을 사용하지 않고 encoder decoder 과정의 반복으로 구성
- Recurrent connection을 제거해 병렬연산 가능

이전 CV + Transformer 노력:

- CNN + SA
- SA for neighborhoods
- SA with small-resolution

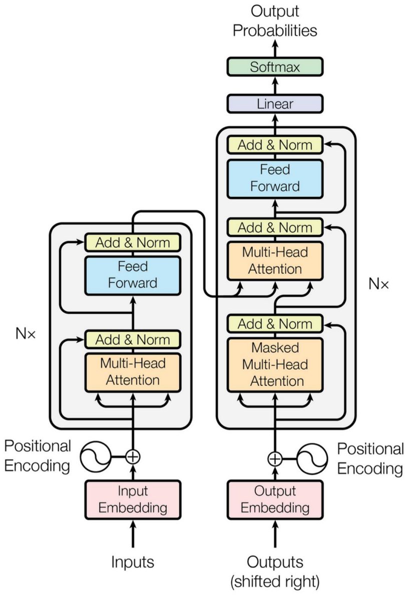
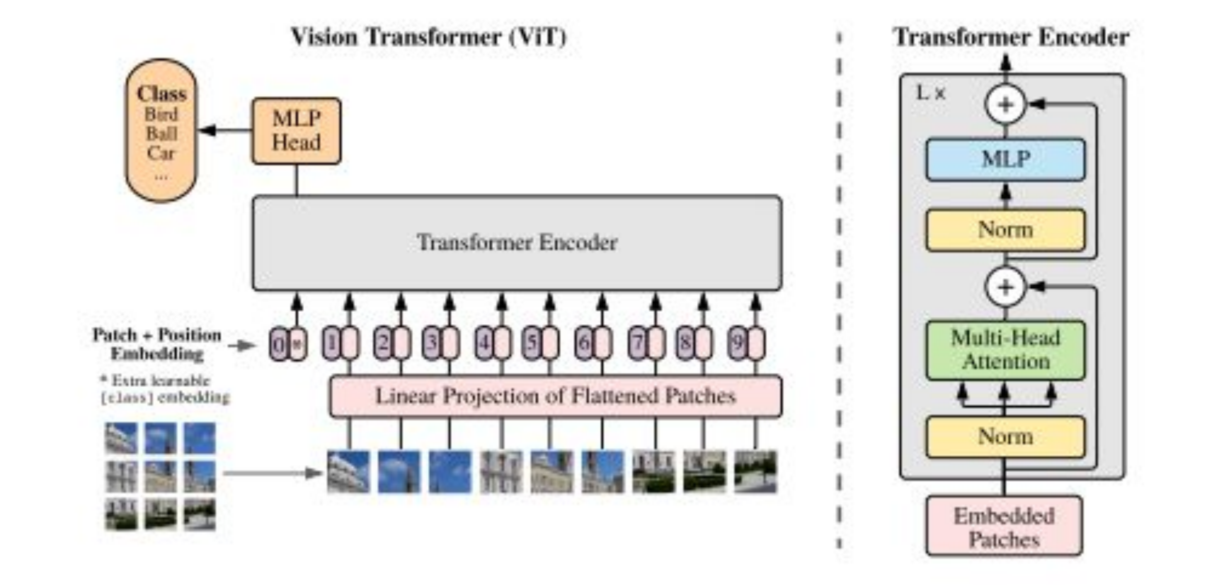


Figure 1: The Transformer - model architecture.

Methodologies

Transformer 구조를 Classification task에 적용

Transformer in CV: token = patch of image

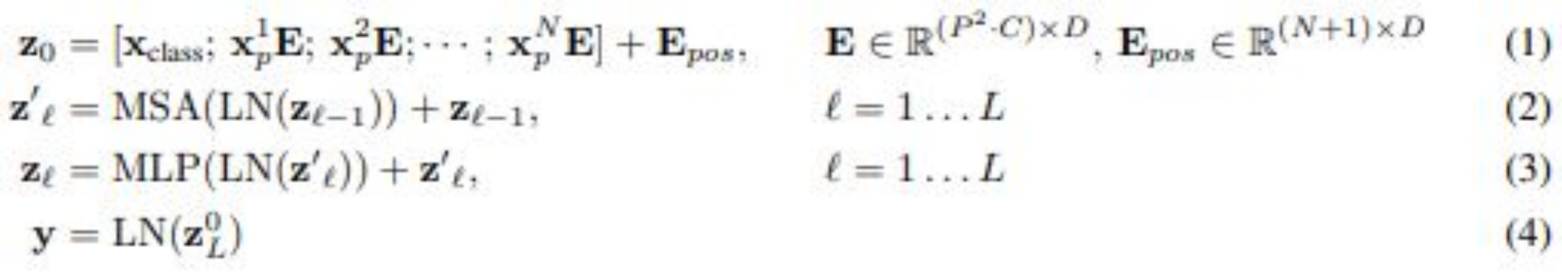


Methodologies 1 - input

Image를 고정된 크기(PxP) 의 patch N개로 분해 (X(HxWxC) → X (PxPxC))

각 patch는 flatten 된 후 D차원 공간으로 linear projected, (i.e., patch embedding)

Learnable embedding을 사용해서 클래스 토큰 prepend



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

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

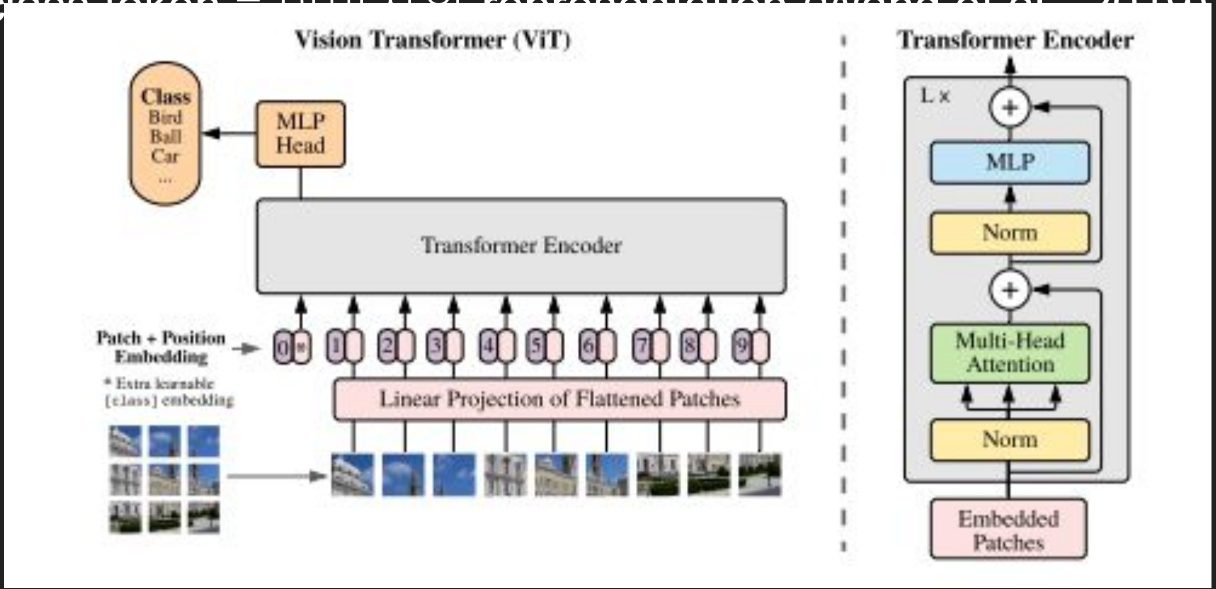
$$\mathbf{z}_\ell = \mathrm{MLP}(\mathrm{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \qquad \ell = 1 \ldots L \qquad (3)$$

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

Methodologies 2 - encoder

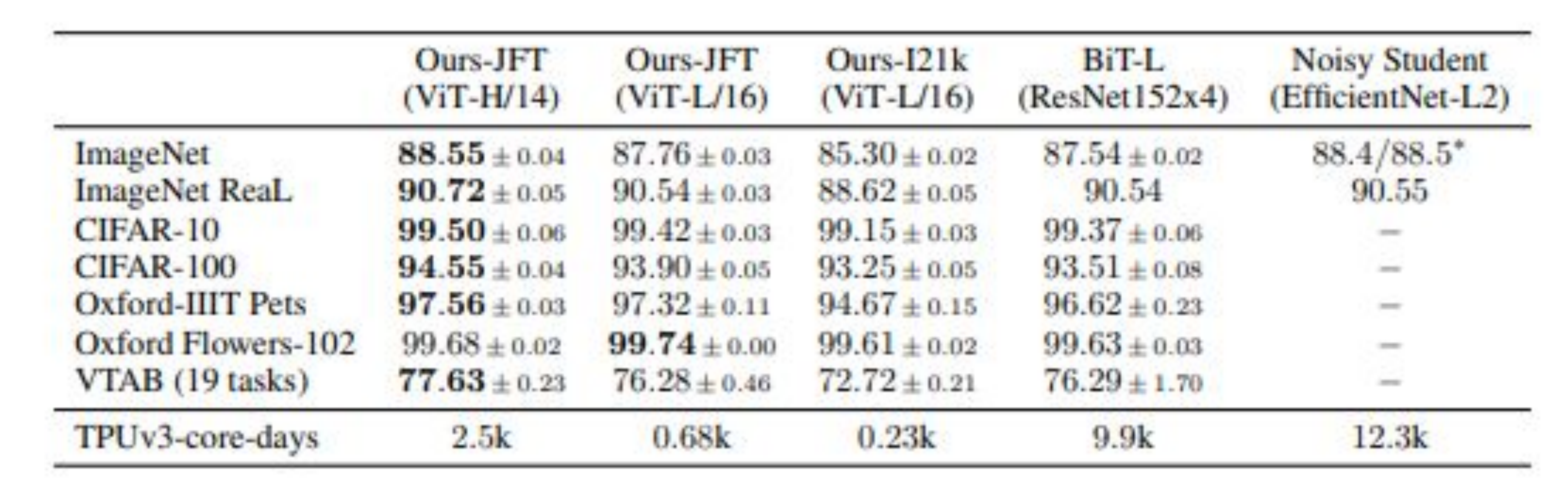
거의 구조가 원 transformer과 유사하나 normalize & residual connection의 위치 변경

L 개의 인코더를 거친 후 class token = 이미지의 representation (Wang et al., 2019)



Experiments

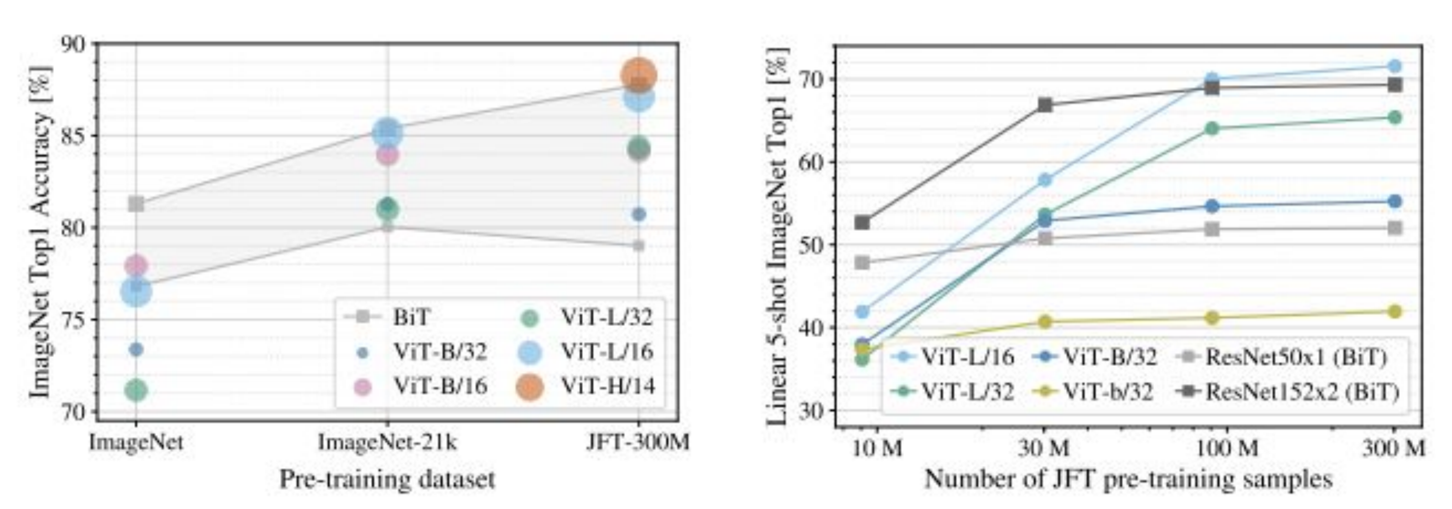
Setup: ResNet / ViT / Hybrid



	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

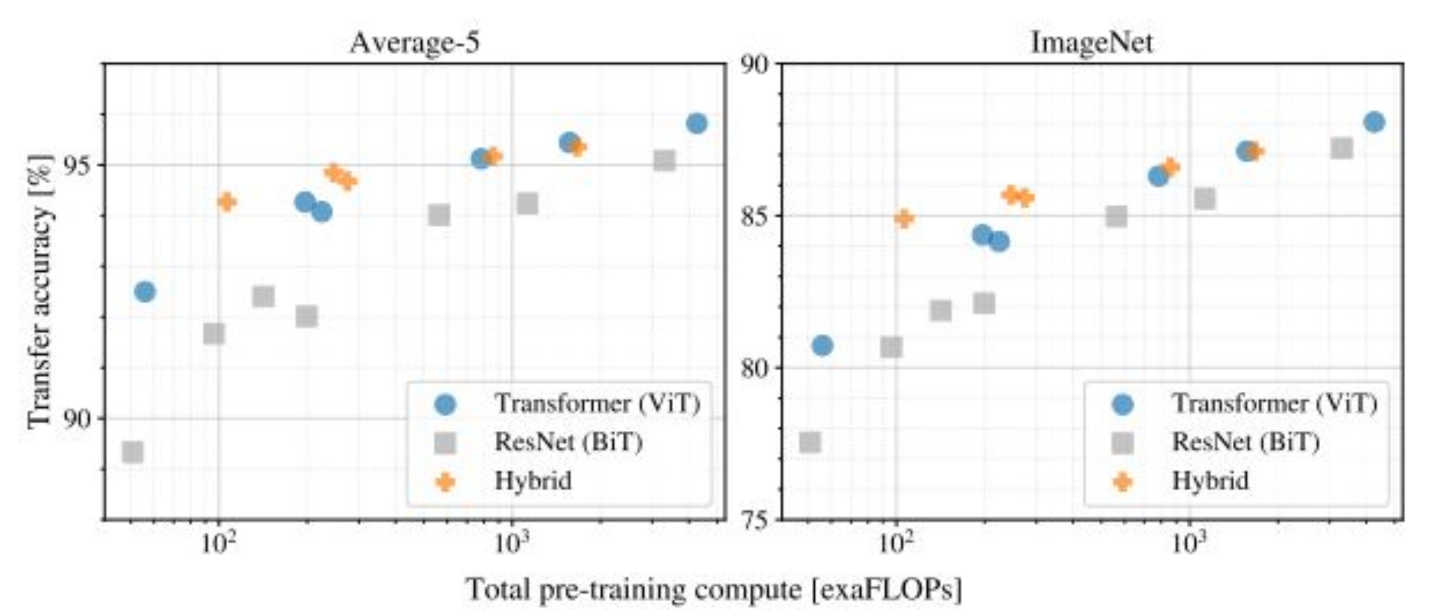
Experiments

Setup: ResNet / ViT / Hybrid



Experiments

Setup: ResNet / ViT / Hybrid

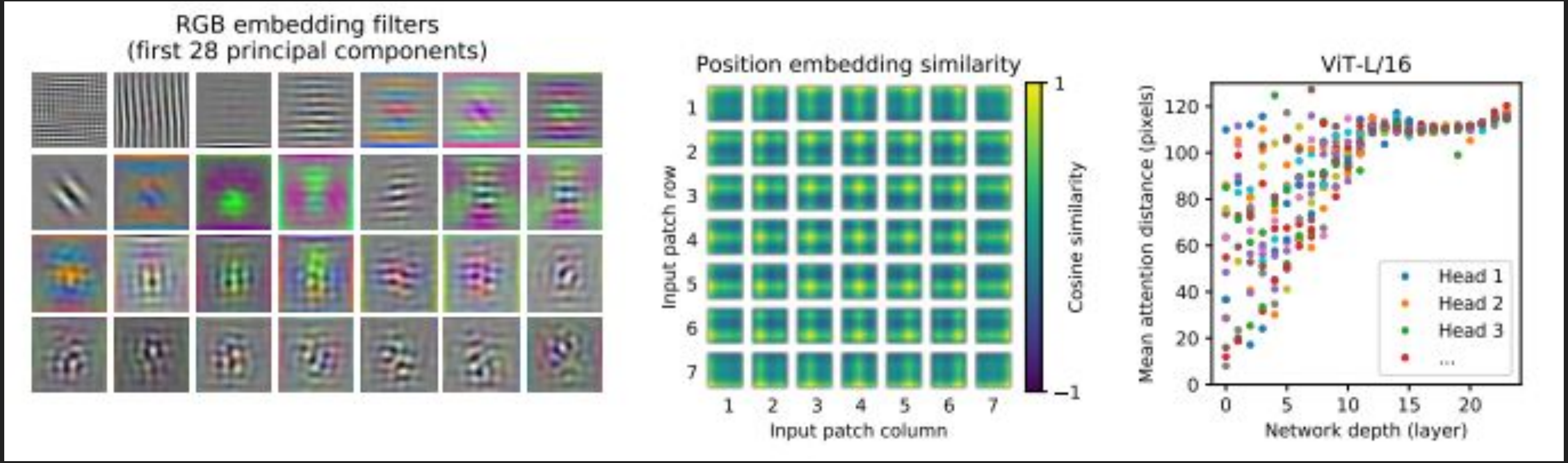


Insights

- **Local vs. Global Attention:** CNNs capture local spatial features through convolutions, whereas ViTs capture global relationships using self-attention.
- **Inductive Biases:** CNNs have built-in inductive biases, such as locality and translation invariance, which make them effective at image tasks with smaller datasets. Vision Transformers, on the other hand, rely on data to learn these patterns, making them more flexible but data-hungry.
- **Training Data Requirements:** Vision Transformers typically require large amounts of training data to achieve their best performance, while CNNs can perform well even with smaller datasets.

Experiments

Setup: ResNet / ViT / Hybrid



Conclusion

- Transformer을 CV task에 적용하는 방법인 vision transformer을 제안
- Image-specific inductive bias를 도입하지 않고 image를 patch의 sequence로 해석함
- 크기가 큰 데이터셋을 활용하면 적은 학습 비용으로 CNN 모델을 능가