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

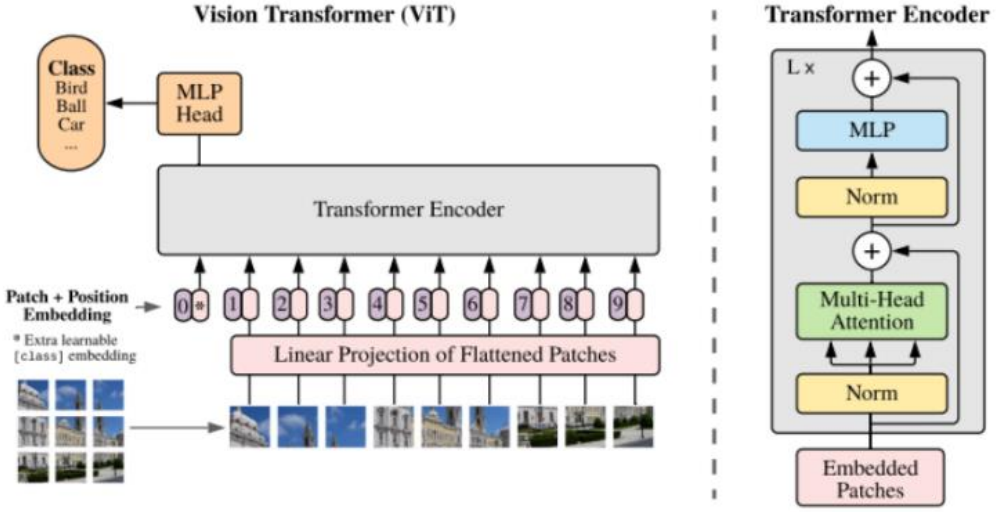
0. Introduction

Transformer architecture는 자연어 처리 작업에서 표준으로 채택되었지만, 컴퓨터 비전 분야에서의 응용은 제한적이다. 그래서 주로 비전 분야에서는 convolution network와 함께 attention을 사용하거나 convolution network의 일부를 대체하는 데 attention을 사용하면서 전체 구조를 유지하는 방식이 사용되었다. 하지만 이 논문에서는 convolution network에 대한 의존성이 필요하지 않으며, 이미지 패치 시퀀스에 직접 적용되는 순수한 트랜스포머가 이미지 분류 작업에서 매우 효과적으로 수행될 수 있음을 보여줄 것이다. 즉, Transformer의 계산 효율성과 scalability를 비전에 활용하고자 한다.

Transformer는 CNN 계열의 모델은 모델링할 수 있는 지역성과 입출력 위치의 동일성 등 이미지 이해에 필수적인 inductive bias를 학습하지 못한다. (Transformer는 CNN보다 상대적으로 inductive bias가 낮다.) 그러나 모델이 더 큰 데이터셋에서 훈련된다면 상황이 달라진다. 본 논문은 대규모 훈련이 inductive bias를 능가한다는 것을 발견했다.

1. Related Work: Vision Transformer의 작동 과정

1단계, 이미지가 있을 때, 이미지를 (P*P) 크기의 패치 N개로 분할하여 패치 sequence x_p를 구축.
2단계, Trainable linear projection을 통해 x_p의 각 패치를 flatten한 벡터를 D차원으로 변환한 후, 이를 패치 embedding으로 사용.
3단계, Learnable class embedding과 패치 embedding에 learnable position embedding을 더함.
4단계, embedding을 vanilla Transformer encoder에 input으로 넣어 마지막 layer에서 class embedding에 대한 output인 image representation을 도출.
5단계, MLP에 image representation을 input으로 넣어 이미지의 class를 분류.



2. Proposed Method: Vision Transformer

Positional Embedding: Vision Transformer에서는 아래 4가지 embedding을 시도한 후, 최종적으로 가장 효과가 좋은 1D position embedding을 사용한다.

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

아래는 전체 구조에 대한 formulation이다. Transformer encoder는 multiheaded self-attention(MSA)와 MLP 블록을 번갈아 쌓은 레이어로 구성된다. 각 블록 이전에 레이어 정규화(LN)가 적용되고, 각 블록 이후에 residual connections이 적용된다.

$$
\begin{aligned}
\mathbf{z}_0 &= [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, && \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D},\ \mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D} \\
\mathbf{z}'_\ell &= \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, && \ell = 1 \ldots L \\
\mathbf{z}_\ell &= \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, && \ell = 1 \ldots L \\
\mathbf{y} &= \text{LN}(\mathbf{z}_L^0)
\end{aligned}
$$

Hybrid Architecture: 혼합 모델에서는 입력 시퀀스를 원본 이미지 패치 대신 CNN feature map에서 형성할 수 있고, 패치 embedding projection을 이 feature map에서 추출된 패치에 적용한다. 특별한 경우로 패치는 공간 크기가 (1*1)일 수 있으며, 이 경우 입력 시퀀스는 feature map의 공간 차원을 평평하게 만들고 Transformer 차원으로 투영하여 얻을 수 있다. 분류 input embedding과 position embedding은 기존 모델과 동일하게 적용한다.

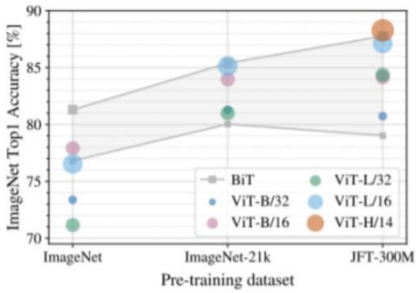
3. Experiments

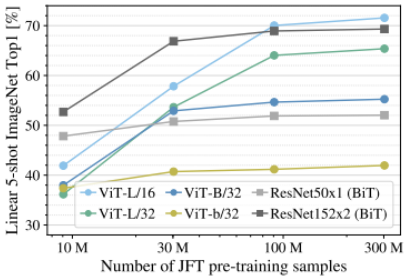
Resnet, Vit 및 Hybrid의 representation learning capability를 평가했다. 각 모델의 데이터 요구사항을 이해하기 위해 다양한 크기의 dataset을 사전 학습사고 많은 벤치 마크에서 작업을 수행한다. 모델의 cost of pre-training을 고려할 때, Vit는 매우 우수한 성과를 보이며, 대부분의 인식 벤치마크에서 SOTA 성능을 더 낮은 pre-training cost로 달성한다. 본 논문은 downstream datasets의 결과를 few-shot 또는 fine-tuning 정확도를 통해 본다. 아래는 인기 있는 이미지 분류 벤치마크에서 SOTA와의 비교이다.

	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

JFT dataset에서 pre-training한 Vit-L/16 모델이 모든 downstream task에 대하여 Bit-L보다 높은 성능을 도출한다.

또한, 오른쪽의 결과는 pre-training dataset의 크기에 따른 fine-tuning 성능을 확인한 것이다. 이는 Vit 모델 크기와 dataset 크기 간의 상관관계를 시각적으로 보여준다. 각 dataset에 대하여 pre-training한 Vit를 ImageNet에 transfer learning한 정확도를 확인한 결과, dataset 크기가 커짐에 따라 Vit 모델의 성능이 향상되고 더 큰 모델이 더 작은 모델을 앞서게 됨을 알 수 있다. 즉, 데이터가 클수록 Vit가 Bit보다 성능이 좋고, 크기가 큰 Vit모델이 효과가 있다고 말할 수 있다.





이 결과는 Vit이 더 작은 dataset에서는 비슷한 computational cost을 가진 Resnet보다 overfit 현상이 더 많이 나타남을 보여준다. 즉, convolutional inductive bias는 작은 dataset에서는 유용하지만, 큰 dataset에서는 데이터로부터 직접 관련 패턴을 학습하는 것이 충분하고 심지어 유익하다는 직관을 강화한다.

위 실험에서 도출된 ImageNet에서의 few-shot 결과 및 VTAB에서의 낮은 데이터 결과는 low-data 전송에도 유망한 성능을 보인다는 것을 확인할 수 있다. 그래서 Vit의 few-shot 특성에 대한 추가적인 분석을 하는 것이 미래 연구의 흥미로운 방향일 수 있겠다고 생각했다.

4. Conclusion

본 논문은 Transformer를 이미지 인식에 직접 적용하는 연구를 수행했다. 컴퓨터 비전에서 Self-attention을 사용하는 이전 연구들과는 달리, 초기 패치 추출 단계를 제외한 아키텍처에 이미지 특정 inductive biases를 도입하지 않았다. 대신 이미지를 패치의 시퀀스로 해석하고 이를 NLP에서 사용되는 표준 Transformer 인코더로 처리했다. 이러한 간단하면서 확장 가능한 전략은 대규모 데이터셋에서의 사전 훈련과 결합될 때 놀라운 결과를 보여준다. 따라서 Vit는 많은 이미지 분류 dataset에서 SOTA를 능가하며, pre-train 비용이 굉장히 저렴하다는 장점을 가진다.

하지만 아직 많은 도전 과제들이 남아있다. 하나는 Vit를 인식 외에 감지 및 분할과 같은 다른 컴퓨터 비전 작업에도 적용하는 것이다. 또 다른 도전 과제는 자기 지도 사전 훈련 방법을 계속 탐구하는 것이다. 초기 실험은 자기 지도 사전 훈련에서의 개선을 보여주지만, 여전히 자기 지도와 대규모 지도 사전 훈련 사이에 큰 격차가 있다. 모델 구조 개선을 통한 성능 향상의 여지가 존재하므로 추가적인 확장을 통한 성능 향상을 기대해본다.

[References]

· https://arxiv.org/abs/2010.11929

· https://www.youtube.com/watch?v=0kgDve_vC1o

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

· https://jeonsworld.github.io/vision/vit/

· https://eehoeskrap.tistory.com/486