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

트랜스포머 구조는 NLP과정의 토대가 된 반면, CV에서의 적용은 한계가 있다. 비전에서 어텐션은 Conv와 함께 적용되거나 전반적인 구조를 유지하며 일부 Conv를 어텐션으로 대체하며 적용한다. 우리는 CNN에 대한 의존이 필요 없음을 보이며 순수 트랜스포머 구조가 이미지 분류에 이용되어짐을 보인다. 많은 데이터를 통해 사전 학습한다면 VIT(Vision Transformer)는 다른 CNN 구조에 비해서 적은 연산량을 통해서 좋은 성능을 나타낸다.

INTRODUCTION

트랜스포머에서 self-attention기반 구조는 NLP에서의 모델이 되었다. 지배적인 접근은 많은 양의 사전 학습후 작은 테스크로의 파인 튜닝을 진행하는 것이다. 트랜스포머의 효율성과 확장성으로 인해 100B 파라미터 넘게 학습이 가능했습니다. 모델과 데이터가 증가함에 따라서, 계속해서 학습이 진행됩니다.

그러나, CV에서 CNN구조가 여전히 우세한 구조입니다. NLP에서의 성공으로 인해 CNN과 self-attention의 결합에 대한 많은 시도가 존재했습니다. 그러나 여러 실패 사례가 존재하였고, 그래서 아직까지도 ResNet구조가 sota를 달성하고 있습니다.

우리는 최소한의 수정을 통해 트랜스포머를 이미지에 직접 적용하고자 합니다. 이를 위해, 우리는 이미지를 패치로 자른 다음 이를 선형 임베딩의 시퀀스로 바꿔 트랜스포머의 주입합니다. 이미지 패치는 NLP에서의 토큰과 같이 다뤄집니다. 우리는 이 모델을 이미지 분류를 위해 학습합니다.

강한 규제 없이 ImageNet과 같은 중간 크기의 데이터셋을 학습시키면 Resnet보다 약간 낮은 성능을 나타냅니다. 이는 트랜스포머는 CNN에 있는 유도적인 편향의 부족에 대한 결함이 있음을 보입니다. 그러므로 불충분한 데이터로는 일반화 학습이 잘 되지 않습니다.

RELATED WORK

트랜스포머는 기계 번역으로 제시되었으며 많은 NLP테스크에서의 Sota 방법입니다. 대용량의 트랜스포머 기반 모델은 대용량 언어뭉치를 통해 학습되어지며 특정 테스크를 위해 파인튜닝되어집니다.

이미지에 셀프 어텐션을 적용하기 위해 각각의 픽셀은 모든 픽셀과 관계를 생각해야합니다. 픽셀의 수에서 이진법적인 비용을 생각하면 이는 현실적인 인풋 사이즈가 아닙니다. 그래서, 트랜스포머 적용을 위한 이미지 처리 방식은 과거에 여러 시도가 있었습니다. 셀프 어텐션을 전체 픽셀에 적용하는 것이 아닌 근처 픽셀간 적용하는 방법, 전체 셀프 어텐션을 위해 근사치를 적용하는 방법, 다양한 크기의 블락을 이용하여 진행하는 방법 등등.. 이 있었습니다. 이러한 연구는 CV task에서 트랜스포머에 대한 비전을 보여주지만, 이전에 하드웨어를 위해 여러 전처리가 필요함을 보여줍니다.

우리와 가장 관련된 여구는 이미지에서 2*2사이즈의 패치를 추출하여 self-attention을 적용하는 것입니다. 이 모델은 ViT와 매우 유사하나, 우리는 대용량의 사전학습된 트랜스포머가 CNN 구조를 이기는가에 대해 밝힘에 목적이 있습니다. 게다가, 2*2방식은, 작은 해상도의 이미지에만 적용이 가능합니다. 그러나 우리는 중간 해상도에도 적용이 가능합니다.

또한, CNN과 attention의 결합에 대한 많은 연구도 존재합니다.

또 다른 관련 연구는 image GPT입니다. 이는 이미지의 해상도와 컬러 공간을 감소시킨 다음 적용합니다. 모델은 생성 모델로 패션에 대해 비지도 학습으로 학습합니다. 그리고 결과는 파인 튜닝되거나 분류를 위해 선형적으로 진행됩니다. 이미지넷에서 72%의 정확도를 기록합니다.

우리의 작업은 이미지 넷 데이터셋보다 큰 데이터를 수집하는 과정이 요구됩니다. 추가적인 데이터 출처는 확인할 수 있으며 우리는 ImageNet-21k와 JFT-300M 두가지 데이터 셋에 초점을 맞춥니다.

METHOD

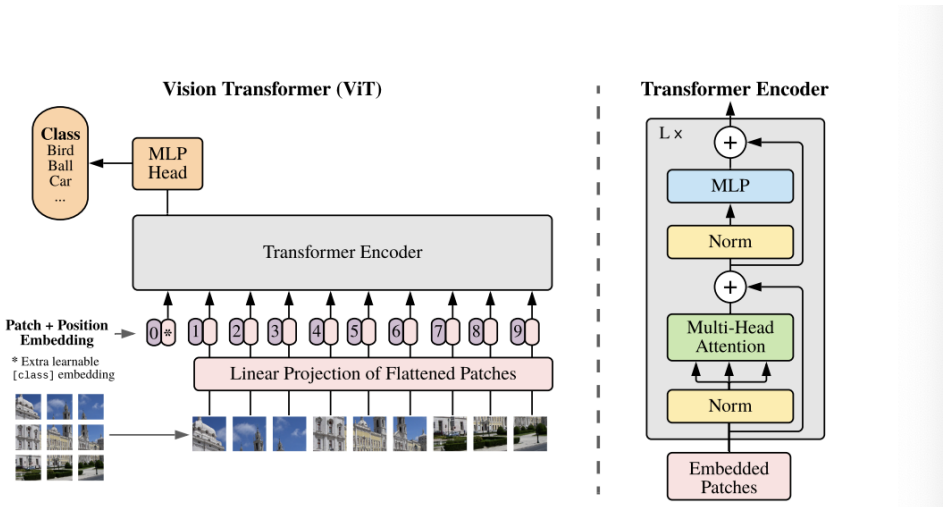

Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification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

우리는 기본적인 트랜스포머 구조를 가능한한 따라갑니다. 이렇게 간단한 구조의 장점은 NLP transformer의 구조와 다양한 효율적인 방식을 우리도 사용 가능하다는 것입니다.

VISION TRANSFORMER

우리 모델은 위와 같이 생겼습니다. 기본적인 트랜스포머는 1차원의 시퀀스 임베딩을 받습니다. 2차원의 미지를 다루기 위해서 우리는 이미지 R(H*W*C)를 펼쳐 놓은 R(N * P^2 * C)로 변형 시킵니다. P는 패치를 뜻하며, N은 채널 당 패치의 수를 말합니다. 트랜스포머는 특정 크기의 차원을 인풋으로 받기 때문에 우리는 패치 시퀀스를 linear를 통해서 고정된 D차원으로 변경해줍니다.

BERT의 class 토큰과 같이 우리는 학습할 수 있는 임베딩을 임베딩 패치 시퀀스에 미리 준비해둡니다. 분류기는 프리트레이닝간 한개의 hidden layer를 통해 진행되어지고 파인 튜닝 간에는 한개의 linear layer로 진행됩니다.

포지션 임베딩 또한 패치 임베딩에 추가되어집니다. 우리는 2D차원의 포지션 임베딩을 주입했을 때 효과를 보지 못했기 때문에 1차원의 포지션 임베딩을 사용합니다.

트랜스포머 인코더는 멀티헤드 셀프 어텐션과 MLP block, Layer norm으로 구성되어 졌으며 모든 블럭에 적용 전 Residual connection이 사용되어집니다.

Inductive bias

우리는 비젼 트랜스포머가 CNN 보다 더 적은 편향을 가지고 있다 말했습니다. CNN에서는 지역적으로 이차원의 구조를 가지고 있으며 전체 모델에 걸쳐 존재합니다. ViT는 오직 MLP레이어에서 지역적이며, self-attention에서는 글로벌입니다. 이차원의 구조는 매우 드뭅니다. 모델 초기에 이미지를 패치 단위로 자르며 파인 튜닝 시 포지션 임베딩을 조정합니다. 반면, 포지션 임베딩은 초기에는 2차원에 대한 아무런 정보를 가지고 오지 않고 처음부터 학습을 진행해야 합니다.

Hybrid Architecture

이미지 패치를 대체하기 위해서 인풋 시퀀스는 CNN의 피쳐 맵으로부터 제작이 가능합니다. 이 모델에서는 CNN 피쳐 맵에서 패치 임베딩을 합니다. 이러한 특히 상황에서 패치는 1*1 크기를 가질 수 있습니다.

FINE-TUNING AND HIGHER RESOLUTION

전형적으로 우리는 ViT를 대규모 데이터 셋에서 사전학습하며 작은 테스크에 대해 파인 튜닝을 진행합니다. 이러한 것을 통해 우리는 사전 학습된 머리를 제거하고 D*K feedforward layer를 부착합니다. k는 작은 테스크의 클래스 개수입니다. 이는 프리트레이닝보다 높은 해상도에서의 파인튜닝에 이점을 줍니다. 높은 해상도의 이미지를 줄 때, 우리는 같은 크기의 패치를 유지하고 큰 효과적인 시퀀스 길이에서 결과를 냅니다. ViT는 시퀀스 길이를 조정할 수 있는 반면, 학습된 position embedding은 더 이상 효과적이지 않습니다. 그러므로 우리는 미리 학습된 포지션 임베딩의 보간을 통해서 적용합니다. 이러한 해상도 조정과 패치 추출은 트랜스포머에 바이어스가 수동적으로 주입되는 유일한 부분입니다.

EXPERIMENTS

우리는 ResNet, ViT, hybrid에 대해 학습 수용성을 평가합니다. 각각의 모델의 데이터 요구를 이해하기 위해 우리는 다양한 크기의 데이터 셋에 대해 사전학습하고 많은 벤지마크 테스크에 대해 평가합니다. 사전 학습 모델의 비용적 측면을 고려하여 ViT는 매우 우수함을 보여줍니다.

	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).

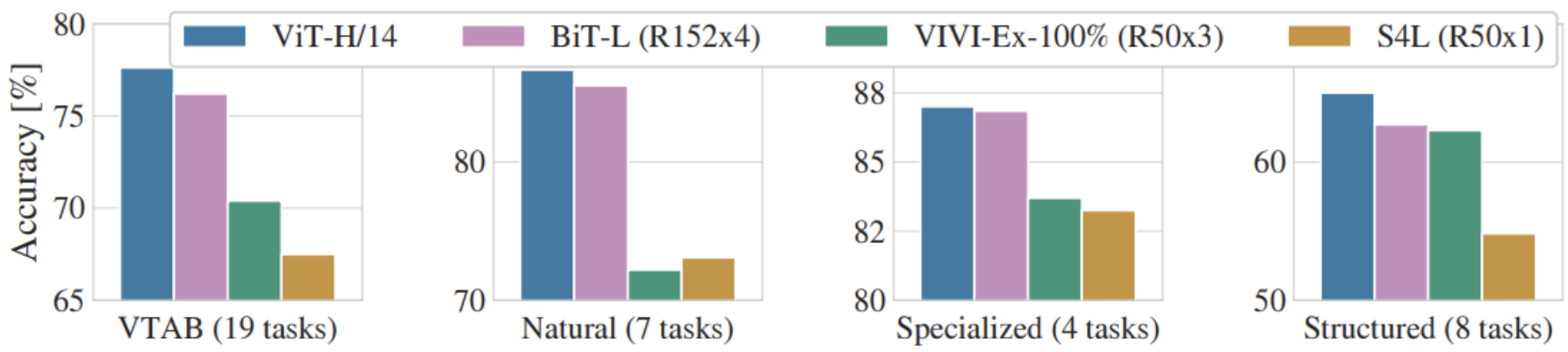

Figure 2: Breakdown of VTAB performance in *Natural*, *Specialized*, and *Structured* task groups.

우리의 ViT model이 좋은 성능을 나타냄을 보여주며 CNN 기반의 모델의 성능을 넘어선 것을 확인할 수 있습니다.

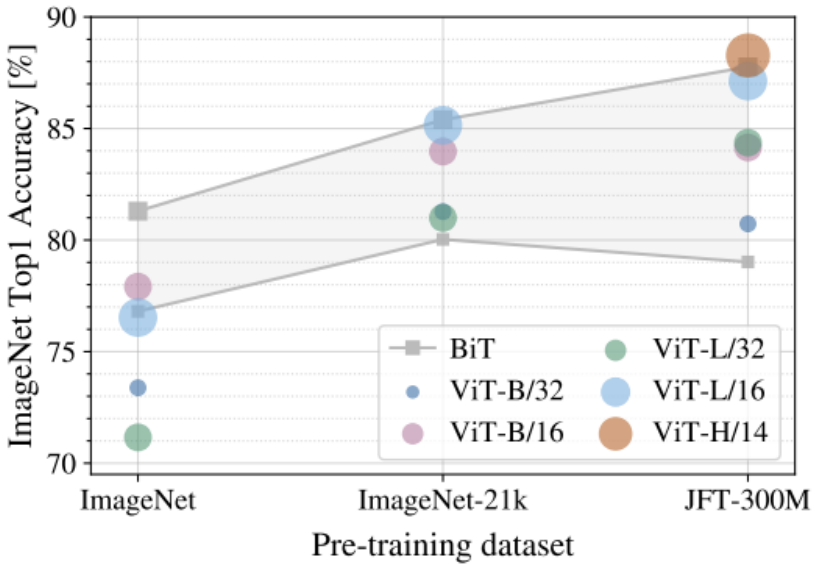

Figure 3: Transfer to ImageNet. While large ViT models perform worse than BiT ResNets (shaded area) when pre-trained on small datasets, they shine when pre-trained on larger datasets. Similarly, larger ViT variants overtake smaller ones as the dataset grows.

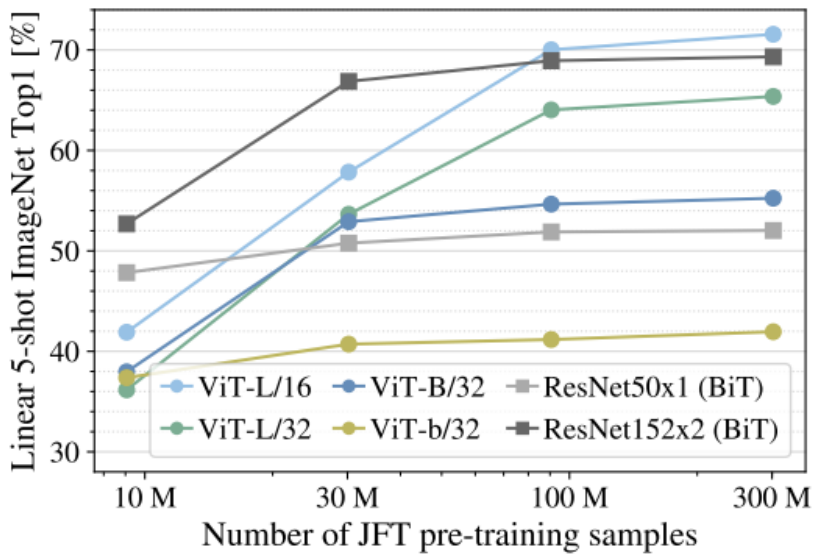

Figure 4: Linear few-shot evaluation on ImageNet versus pre-training size. ResNets perform better with smaller pre-training datasets but plateau sooner than ViT, which performs better with larger pre-training. ViT-b is ViT-B with all hidden dimensions halved.

사전 학습 데이터 수량에 따른 성능을 확인하면 데이터 셋의 수가 적은 ImageNet의 경우 ViT의 성능이 ResNet보다 낮으며 데이터 셋이 늘어날 수록 ViT가 역전하는 모습이 보입니다.

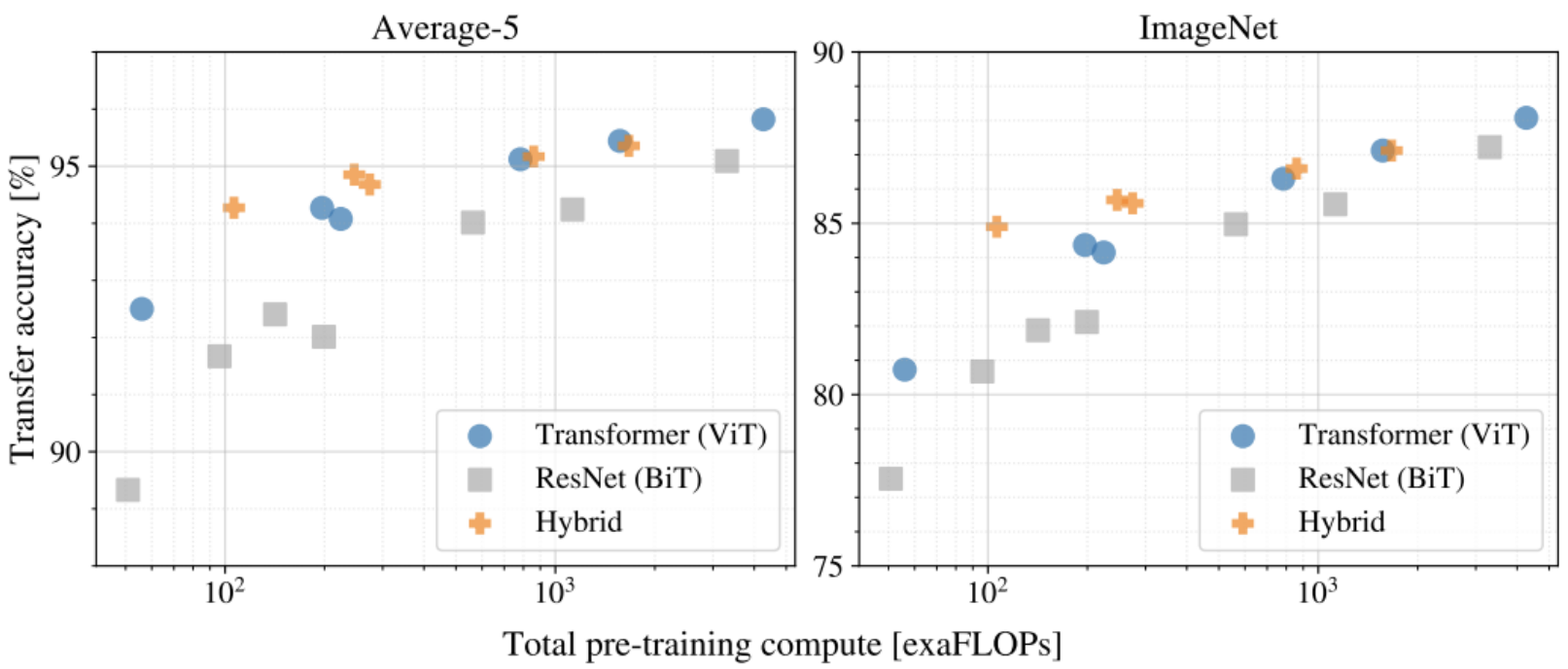

Figure 5: Performance versus pre-training compute for different architectures: Vision Transformers, ResNets, and hybrids. Vision Transformers generally outperform ResNets with the same computational budget. Hybrids improve upon pure Transformers for smaller model sizes, but the gap vanishes for larger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

사전 학습 비용에 따른 성능입니다. 사전 학습 비용이 같을 때 ViT의 성능이 ResNet보다 높으며 Hybrid는 대체로 성능이 가장 높다가 사전 학습이 많이 이루어질 경우 pure ViT와의 갭이 거의 없어집니다.

CONCLUSION

우리는 이미지 인식에서의 트랜스포머 적용을 연구했다. 이전과는 다르게 셀프 어텐션을 CV에서 사용하며 우리는 이미지 편향을 패치 추출 단계에서 도입하지 않습니다. 대신 우리는 NLP처럼 이미지를 토큰 시퀀스와 같이 사용합니다. 그리고 대규모 데이터 셋의 사전 학습을 요구하며 이는 상대적으로 싼 비용을 요구합니다. 또한 ViT는 CNN 구조의 모델의 성능을 이겨냈습니다.