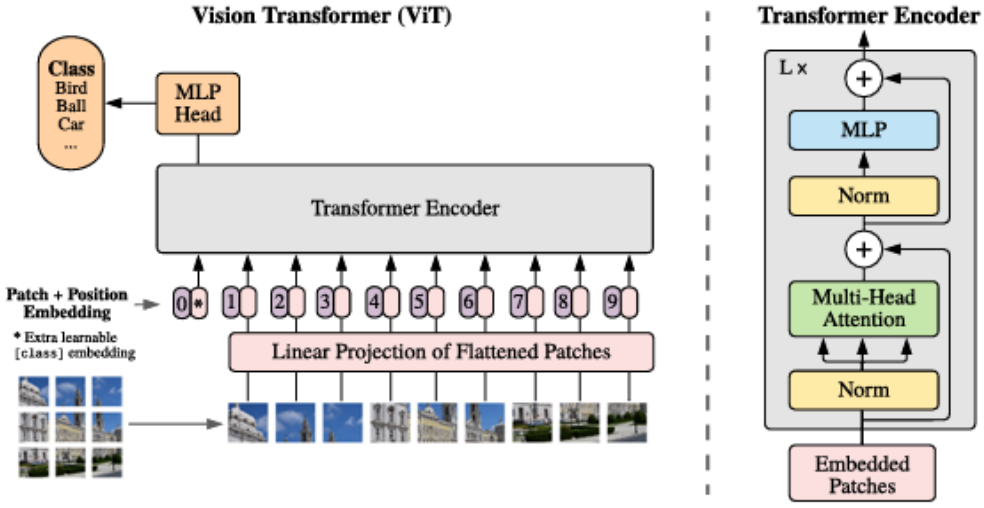
[논문리뷰]Vision Transformer

Introduction

- **Self Attention 기반의 Architecture인 Transformer는 NLP에서 부정할 수 없는 선택지가 되었다. Transformer는 일반적으로 거대한 Text Corpus를 바탕으로 학습을 진행한 뒤, 특정 Task를 수행하기 위하여 Specific한 Dataset을 통해 Fine Tuning과 함께 사용된다.**
- **이러한 Transformer 구조의 계산적 효율성과 높은 확장성으로 인해 지금까지도 계속 Parameter 수가 늘어난 거대한 모델을 보다 거대해진 Dataset을 통해 학습함으로써 계속적으로 성능이 개선되고 있으며, 아직까지는 성능 지표의 발전에 있어서 Saturation이 나타나지 않고 있다.**
- **반면, Computer Vision 영역에서는 여전히 CNN이 우세한 상황이다.**
- **본 논문에서는, Transformer의 NLP에서의 큰 선전에 영감을 받아서 표준 Transformer 구조를 이미지에서 직접적으로 활용할 수 있게끔 최소한의 수정을 통해 변경하였다.**
- **해당 방식은 하나의 Image를 작은 Patch로 나누고, 이러한 Patch들의 Sequence를 Transformer에 전달하는 방식으로 이루어진다. 각각의 Patch들은 마치 NLP에서의 Token(word)와 같이 처리되며 본 논문에서는 일반적인 지도 학습의 방식으로 이미지 분류를 학습시켰다.**
- **실험 결과, ImageNet에서는 비슷한 크기의 ResNet과 비교했을 때 정확도가 많이 떨어지는 모습을 보였다. 이러한 현상은 모두 예측 가능한 것으로, Transformer는 CNN에서 활용되는 Locality 정보와 Translation Equivariance등의 Inductive Bias가 존재하지 않기 때문에 충분하지 않은 데이터에서는 학습이 잘 이뤄지지 않기 때문이다.**
- **반면, 데이터셋의 크기가 커졌을 때(14M~300M), ViT가 이러한 CNN의 Inductive Bias를 능가하는 모습을 보여주었다.**

ViT Method



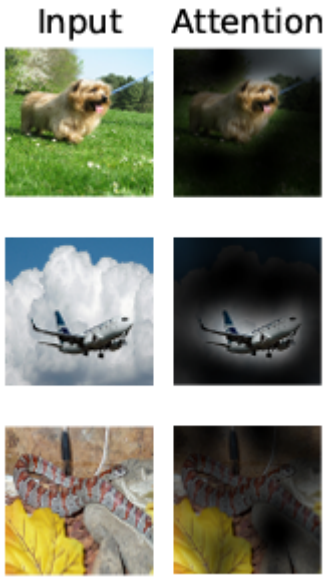
- ViT는 Input으로 Image를 고정된 사이즈의 Patch로 분할한 뒤, 각각의 Patch들을 Linear Projection을 통해 임베딩하고, Sequential Position 정보를 담고 있는 Position Embedding을 추가하여 Transformer의 Encoder에 입력하게 된다.
- 이러한 방식은 단순히 Image를 Sequence 형태로 바꾸는 부분만 추가되었을 뿐, 실제로는 Original Transformer와 거의 유사한 방식으로 동작하게 된다. 이를 통해, 기존에 잘 연구되었던 Transformer Architecture를 사용할 수 있을 뿐만 아니라 효율적인 구현이 가능하도록 했다.
- 단순히 Image Patch로 나누는 방식은 Transformer로 하여금 2D Spartial Information을 매우 제한적으로 사용하게 만들기 때문에 CNN에 비해서 낮은 Inductive Bias를 가지기 때문에 충분히 많은 데이터를 통해 학습을 진행하여야 한다. 본 논문에서는 JFT-300M dataset을 통해 Pretrain을 진행했다.
- 이를 개선하고 Spartial Information을 최대한 활용하게끔 유도하기 위해서, Raw Image Patch 대신, CNN의 Feature Map을 얻어낸 뒤, 이를 Flatten하여 Input Sequence를 만들어낼 수도 있다.
- 일반적인 Transformer의 방식을 그대로 차용하여 Large Dataset에 일단 Pretrain을 진행하고, 그 후 작은 규모의 downstream task dataset에 fine tuning을 진행한다.
- 최근 연구들의 결과에 따라서, Pretrain에 사용됐던 Resolution 보다 더 높은 Resolution을 통해 fine tuning을 진행하여 성능을 더욱 높이고자 하였다.
- 이 때, 해상도를 늘린다고 하여 Patch Size를 키우지는 않고, Patch Size는 항상 Fix 상태로 유지하고, 단순히 Sequence의 길이를 늘려주는 방식으로 해상도를 키운다.
- 하지만 이런식으로 Resolution을 키워주게 되면 Pretrain에서 사용됐던 Position Embedding의 정보가 훼손될 수 있다. 본 논문에서는 이를 방지 하기 위해서, 원본

Image에서의 상대적인 위치를 파악하기 위한 2D interpolation을 진행한다. 저자들은 이것이 2D 구조에 대한 정보를 임의로 집어넣는 유일한 Inductive Bias라고 주장한다.

Result

	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

- 학습 결과, ViT 방식이 충분한 크기의 데이터셋(JFT-300M)으로 학습된 경우 CNN SOTA 모델들과 비교했을 때, 뒤쳐지지 않는 모습을 보였다.
- 또한, 모델을 학습하는데 걸리는 Resource에 대한 지표인 TPUv3-core-days 또한 논문이 발표될 당시 SOTA였던 EfficientNetL2에 비해서 훨씬 작은 값을 가지는 것을 통해 비교적 적은 Resource로도 학습이 가능하다고 주장한다.



- Attention Score가 높은 영역을 확인해보면 실제로 Classification에 유의미한 부분이 Attention Score가 높다는 것을 확인할 수 있으며, 이는 의도한대로 Image를 학습했다는 것을 의미한다.

- **저자들은 대규모 데이터셋을 활용한 Pretraining 방식 외에도, BERT와 같이 Masked Patch Prediction을 사용한 Self Supervised Learning을 시도해보았지만, 대규모 데이터셋에 대한 전이학습에 비해서 성능이 많이 떨어졌다고 한다.**

Conclusion

- **본 논문에서는 NLP에서의 Transformer의 크나큰 활약을 보고 감명을 받아, Vision Task에서도 BERT와 같은 형태의 Transformer Encoder를 활용할 수 있도록 약간의 수정을 진행한 ViT Architecture를 제안하였다.**
- **비록 Transformer는 CNN에 비해서 낮은 Inductive Bias를 가지고 있기 때문에, ImageNet 학습만으로는 CNN의 성능을 이길 수 없었지만, JFT-300M과 같은 거대 데이터셋을 활용한 학습을 통해 CNN의 Inductive Bias를 능가하였다.**
- **뿐만 아니라, CNN 대비 낮은 학습 리소스를 활용하며, SOTA 모델에도 맞먹는 높은 정확도를 가지고 있어 Vision 영역에서의 Transformer의 유망성을 보인 사례라고 할 수 있다.**
- **본 논문에서는 비록 SOTA를 크게 뛰어넘지도 못했고, Object Detection, Segmentation에서의 활용등의 실험을 진행하지는 못했지만 Transformer를 Vision Task에 적용하는 새로운 시도를 진행하였으며, CNN과 비슷한 성능을 얻어냈다는 점에 큰 의의가 있다. 연구분야를 새롭게 개척했다고 해야 하나?**
- **실제로 본 논문이 발표되고 얼마 되지 않아서 DeiT, Swin Transformer, CMT, CvT 등 수많은 방식이 연구되었으며 끊임없는 성능 향상을 나타내고 있다. 최근엔 아예 SOTA Leaderboard를 Transformer류가 점령한 상태이다.**