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

| | |
|---|---|
| 문서 유형 | Transformer + alpha |
| 상태 | 진행 중 |
| 작성자 | |

💡 [ Hongmungwan : Kaggle Study Group] 에서 참여 중인 프로젝트를 위해 공부했던 'ViT(Vision transformer)' 내용에 대해 정리 하고자 한다. 이번 포스팅에서는 'ViT'를 제안한 **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale** 논문 내용을 바탕으로 참고 자료와 함께 다시 정리 하였다.

## 1. Introduction

- **NLP의 Transformer에서 모델 수정을 최소화하면서 표준 Transofrmer를 이미지에 직접 적용**
- **Image를 작은 patch로 분할하고, 이러한 patch를 linear ebedding 의 sequence를 transformer의 input으로 사용**
- mid-size scale dataset에서는 ResNet 등 기존 CNN base model 보다 좋은 성능을 보이지 않음

  **Equivariance or Locality 즉, CNN 고유의 Inductive bais를 고려할수 있는 기능이 transformer에는 없기 때문에 불충분한 양의 데이터셋에서는 일반화가 잘 되지 않는 문제점 존재**

  → large scale dataset으로 pre-train 이후 fine-tuning 할때 좋은 성능을 보임

  (ImageNet, CIFAR-100 benchmark에서 SOTA 달성)

## 2. Related Work

- Image에 대한 self-attention의 단순한 적용은 각 pixel이 다른 모든 pixel들에 attend될 것을 요구
- pixel 수에 대해 quodratic한 복잡도를 가지며 이로인해 현실적인 다양한 input size로의 확장은 어려움
- Transformer를 image processing에 적용하기 위해  approximation 방법들이 시도됨

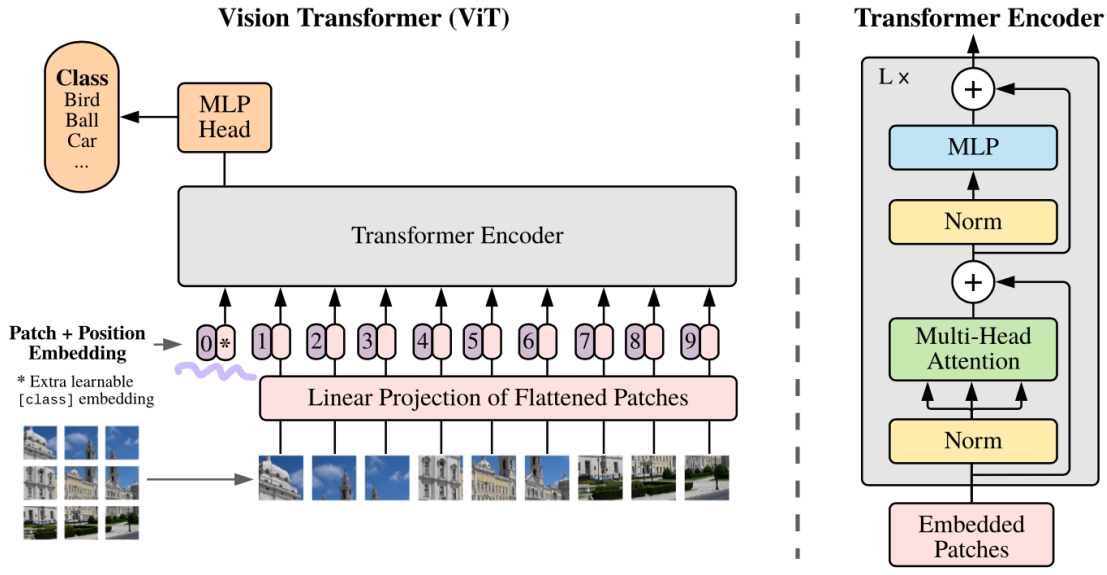
  (local self-attention, sparse attention, applying it in blocks of varying size, etc.)
- CNN과 self-attention을 결합하는 시도의 연구들도 다수
- Image resolution과 color space를 줄인 image pixel들에 transformer를 적용한 Generative model

  → a.k.a iGPT (ViT와 가장 유사한 방법의 연구)

# 3. Method

- model 설계시 original transformer와 최대한 유사하게 구성함

  → 쉽게 확장 가능한 original transformer 구조와 효율적인 구현을 바로 사용할수 있기 때문

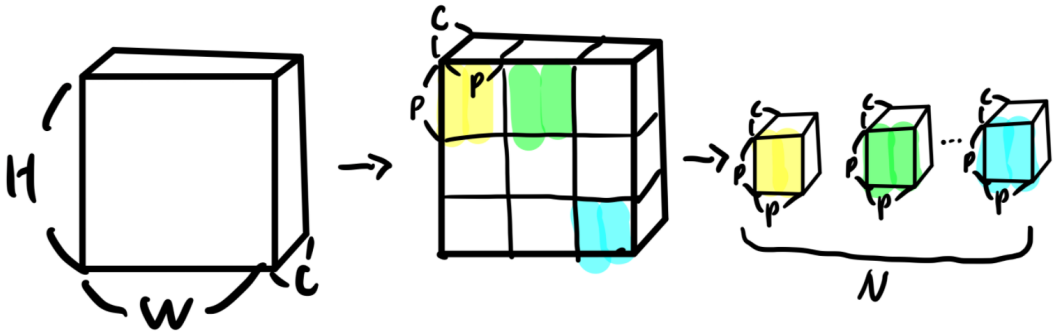



[ 그림 1 ] ViT(Vision Transformer) architecture

▼ **Vision Transformer architecture**

- **Patch Embedding**

- 3차원인 Image($x \in \mathbb{R}^{H \times W \times X}$, (H: Height, W: Width, C: Channel))를 2차원 Image patch($x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$) 로 flatten
- $P$는 각 패치의 가로/세로에 해당하는 크기, $N = HW/P2$ 으로 Image patch 개수 이자 input sequence의 length




[ 그림 2 ] Exemple  Convert to 2D image patch

(** 출처 : Yeongmin's Blog - Vision Transformer Review)

- Transformer 인코더는 batch size, sequence length, hidden size 과 같은 모양의 입력을 사용
- **Image patch($P^2 \cdot C$)를 학습가능한 parameter($E \in \mathbb{R}^{(P^2 \cdot C) \times D}$ 를 이용하여 $D$ 크기의 vector로 linear projection (선형변환)** (D : latent vector size)

- **Classification token**
  - Embedding된 patch 맨 앞에 하나의 학습 가능한 형태의 [ class ] token  벡터를 추가

    (BERT [CLS] token과 같은 기능)
  - 임베딩벡터($z_0^0 = x_{class}$)는 Transformer의 여러 encoder층을 거쳐 **최종 output($z_L^0$)으로 나왔을 때, 이미지에 대한 1차원 representation vector로써의 역할을 수행**

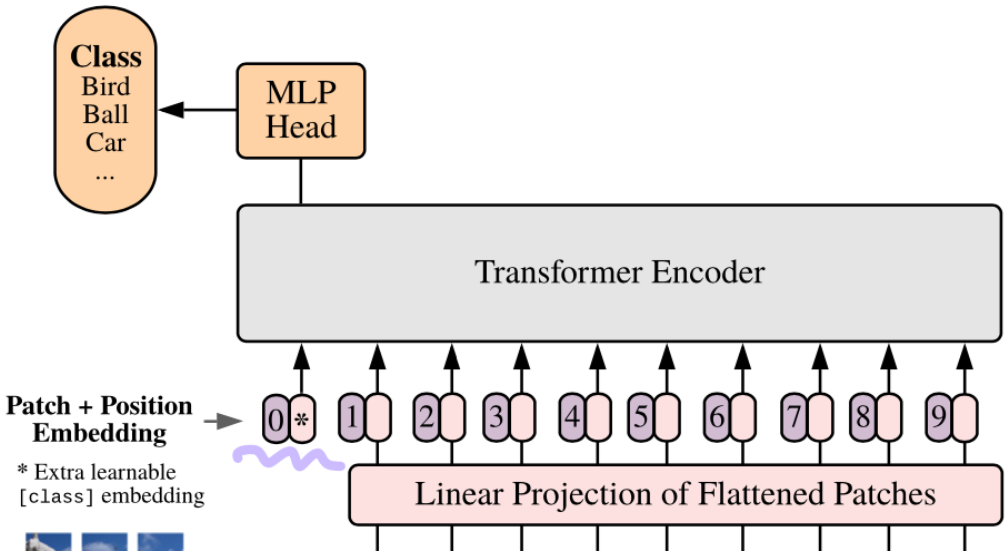
    (sequence의 첫번째 위치에 추가함)
  - image representation vector 위에 classification head를 붙여서 사용하는데, **Pre-train 에서는 하나의 hidden layer를 가지는 MLP로 구현**되고, **fine-tuning에는 단일 linear layer로 구성**
- **Trainable position embeddings**

- 각 패치의 위치 정보를 제공하기 위해 추가적인 각 위치별로 학습 가능한 positional embedding ($E_{pos} \in \mathbb{R}^{(N+1)\times D}$) 사용
- 논문에서 2D image patch의 상대적인 위치를 인코딩 할 수 있는 Advanced positional embed-ding 방식을 시도해봤지만 눈에 띄는 성능 향상을 관찰하지 못했다고 함

- **Transformer encoder**
  - MSA(Multi-head self-attention) layer와, MLP(Multi Layer Perceptron) 가 교차로 입력

  **▼ 모든 block에 LN(Layer normalization) 적용하며, 모든 block이후에 residual connection**

  - $z_0 = [x_{class}; x_p^1 E; x_p^2 E; \ldots; x_p^n E] + E_{pos}$
  - $z_l' = MSA(LN(z_{l-1})) + z_{l-1}$
  - $z_l = MLP(LN(z_l')) + z_l'$
  - $y = LN(z_L^0)$




[ 그림 3 ] Transformer Encoder

**▼ Hybrid architecture**

- CNN feature map을 transformer encoder의 input sequence로 사용하는 방법도 실험 진행

  ( feature map의 결과를 flattening 한 뒤, Trasnformer의 차원으로 projecting )

- 즉 CNN위에 Transformer encoder를 쌓은 구조

## 3.1 Train method

▼ **Fine-tuning**

- large dataset으로 pre-train하고, 이를 downstream task에 fine-tuning
- BERT의 LM을 이용한는 방법과 동일

▼ **higher resolution fine-tuning**

- patch size (*P*×*P*)는 유지하고, Transformer input sequence length(N)를 늘리는 방향
- 어떠한 input sequence length에 대해서 처리 할 수 있으나, Pretrain된 postion embedding을 사용하지 못함 (diff. shape)

  2D interporation 방식으로 원본 이미지 위치에 따라 Pretraied position embedding을 늘림

# 4. Experiments & Result

- ResNet 구조로 supervised transfer learning을 수행한 BiT(Big Transfer), EfficientNet구조로 ImageNet, JFT-300M를 semi-supervised learning한 Noisy Student 두 SoTA 방식을 비교
- **대부분 기존 SoTA를 능가하거나 비슷한 성능을 보이나, 학습 시간의 경우 ViT가 앞도적으로 적게 걸림**

| | 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).

[ 그림 4 ] experiments Benchmark Dataset

- 사전학습 데이터량에 따른 성능 비교 실험 진행
- **데이터의 양이 작을 경우 큰 모델(ViT-L)이 작은 모델(ViT-B)보다 성능이 떨어짐**

(few-shot의 경우도 동일함)

- 즉, 모델 크기의 이점을 완전히 가져가려면 그만큼 많은 데이터가 필요

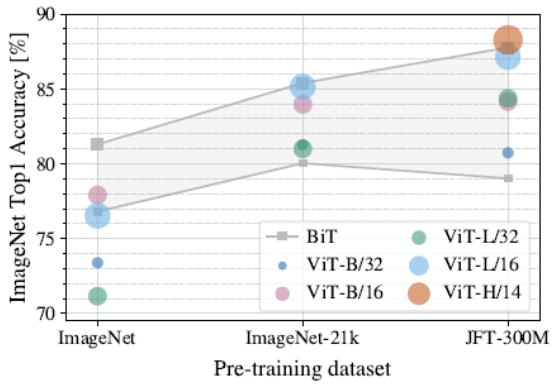



Figure 3: Transfer to ImageNet. While large ViT models perform worse than BiT ResNets (shaded area) when pre-trained on small datasets, they shine when pre-trained on larger datasets. Similarly, larger ViT variants overtake smaller ones as the dataset grows.

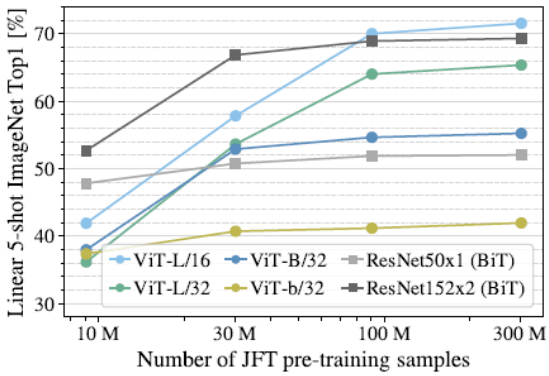



Figure 4: Linear few-shot evaluation on ImageNet versus pre-training size. ResNets perform better with smaller pre-training datasets but plateau sooner than ViT, which performs better with larger pre-training. ViT-b is ViT-B with all hidden dimensions halved.

[ 그림 5 ] Experiments diff. Pretrain Dataset size

# 5. Reference


[1] An Image is Worth 16*16 Words: Transformers for Image Recognition at Scale, ICLR 2021

[2] Vision Transformer Review

[3] Vision Transformer(ViT) 논문리뷰