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

Summarized by: 202355514 강지원

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby

1 Google Research, Brain Team

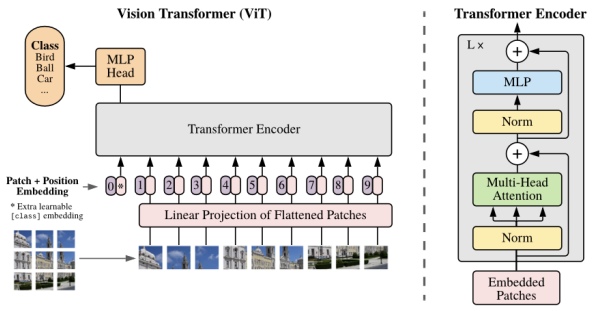


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification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

Transformer는 Self-Attention을 기반으로 한 구조로 자연어처리 분야에서 가장 널리 쓰이는 모델이다. 일반적으로 대규모 텍스트 데이터로 사전학습 한 뒤, 작은 데이터셋으로 fine-tuning 하는 방식이다. Transformer는 계산의 효율성과 확장성 덕에 1000억 개 이상의 파라미터를 가진 포대형 모델까지도 학습 가능하며 성능은 여전히 한계에 다다르지 않고 향상되고 있다.

본 논문에서는 NLP의 Transformer 확장 성공 사례에서 영감을 받아, 작은 이미지 패치들을 단어처럼 처리해 Transformer에 직접 입력하면 대규모 데이터에도 활용 가능하며 cnn을 능가하는 성능을 나타냄을 보인다. 즉 데이터의 규모가 커지면 CNN의 구조적 이점보다 Transformer의 확장성이 더 강력함을 보이고자 한다.

Vision Transformer(ViT)는 이미지를 16×16 크기의 패치로 나누어 각 패치를 벡터로 변환한 뒤, 이를 문장의 단어처럼 Transformer에 입력하는 방식으로 작동한다. 각 패치에는 위치 임베딩이 더해지고, BERT의 [class] 토큰과 유사한 이미지 대표 토큰이 추가되어 전체 이미지를 분류하는 데 활용된다. 모델 구조는 기존 Transformer와 동일하게 다중 자기어텐션(Multi-Head Self-Attention)과 MLP 블록이 교차로 쌓여 있으며, CNN과 달리 지역성이나 이동 불변성과 같은 구조적 편향이 거의 없어 공간적 관계를 학습을 통해 스스로 익혀야 한다. 또한, 원시 패치 대신 CNN의 feature map을 Transformer 입력으로 사용하는 하이브리드 구조도 제안되어 두 접근법의 장점을 결합하였다.

ViT는 대규모 데이터로 사전학습한 후 작은 데이터셋에 맞춰 미세조정하며, 해상도를 높여 학습할수록 성능이 향상된다. 이때 해상도 변화로 인해 기존 위치 임베딩의 의미가 달라질 수 있어, 2차원 보간을 통해 이를 조정하는데, 이는 ViT가 이미지의 2D 구조를 반영하는 거의 유일한 편향 요소이다.

ViT는 ImageNet, ImageNet-21k, JFT-300M 등 다양한 규모의 데이터셋에서 사전학습한 뒤, ImageNet, CIFAR, VTAB 등 여러 벤치마크로 전이학습을 수행하였다. BERT 구조를 기반으로 한 Base · Large · Huge 모델 구성을 사용했으며, ResNet 및 하이브리드 모델과의 비교를 통해 학습 효율과 확장성을 검증하였다. 그 결과, ViT-L/16과 ViT-H/14 모델은 기존 SOTA CNN 모델인 BiT와 Noisy Student를 능가하면서도 훨씬 적은 연산 비용으로 학습되었다. 특히 ViT-H/14는 다양한 벤치마크에서 최고 성능을

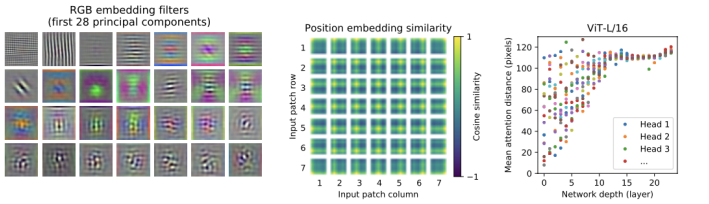


Figure 2: Left: Filters of the initial linear embedding of RGB values of ViT-L/32. Center: Sim- ilarity of position embeddings of ViT-L/32. Tiles show the cosine similarity between the position embedding of the patch with the indicated row and column and the position embeddings of all other patches. Right: Size of attended area by head and network depth. Each dot shows the mean attention distance across images for one of 16 heads at one layer. See Appendix D.7 for details.

기록했고, ViT-L/16 역시 대규모 데이터로 사전학습 시 효율적이고 경쟁력 있는 결과를 보였다.

ViT의 성능은 데이터셋의 규모에 크게 의존하며, 작은 데이터에서는 ResNet보다 과적합이 발생하지만, 충분히 큰 데이터에서는 스스로 시각적 패턴을 학습해 CNN을 능가하였다. 스케일링 실험 결과, ViT는 계산 효율성과 확장성 면에서 뛰어나며, 모델 규모가 커질수록 성능이 지속적으로 향상되었다. 또한 내부 표현 분석을 통해 ViT가 패치 단위로 세부 구조를 학습하고 위치 임베딩을 통해 공간적 관계를 스스로 형성함을 확인했으며, Self-Attention을 통해 전역 정보를 통합하고 깊은 층으로 갈수록 의미 있는 이미지 영역에 더 집중하는 특성을 보여주었다.

마지막으로, BERT의 방식을 적용한 masked patch prediction 기반 self-supervised pre-training 실험에서는 ViT-B/16이 ImageNet에서 79.9%의 정확도를 달성하며 학습 효율 향상을 보였으나, 여전히 지도학습과의 성능 격차가 남아 있었다. 이는 향후 contrastive learning 기반의 자가지도 학습이 ViT 발전의 중요한 방향으로 제시하였다

결론적으로 본 연구는 이미지를 패치 단위로 분할해 표준 Transformer 인코더로 처리하는 단순한 구조의 ViT를 제안하였으며, 대규모 사전학습을 통해 기존 CNN 수준의 성능을 효율적으로 달성했다. 향후 과제로는 탐지 · 분할 등 다른 비전 과제 적용, self-supervised pre-training의 고도화, 그리고 모델의 추가 확장이 제시되었다.

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. *"An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale."* International Conference on Learning Representations (ICLR), 2021.

[2] *"Vision Transformer: A Visual Guide to Vision Transformers."* *PyTorch KR* , 2023. https://discuss.pytorch.kr/t/vision-transformer-a-visual-guide-to-vision-transfor 4158