# Vision Transformer

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

---

### ▼ 1. 어떤 문제를 풀고자 했는가? (Abstract)

💡 Transformer in CV
: attention이 convolution network와 함께 적용되거나, 특정 요소를 대체하기 위해 사용되었음!
→ CNN 구조 대부분을 transformer로 대체함으로써 더 좋은 이미지 분류 성능!
→ but, 이때 충분한 데이터 셋으로 pre-train하고, target task에 fine-tuning을 해야 성능을 제대로 발휘할 수 있음!

### ▼ 2. 어떤 동기/상황/문제점에서 이 연구가 시작되었는가? (Introduction)

💡 **기존 연구**
- CV에서는 Convolutional 구조가 우세하게 사용됨!
- self-attention과 결합하려는 시도 있었지만, 현대 하드웨어에서는 효과적으로 scaling 불가

In This 논문…
- standard transformer를 최소한의 수정으로 직접 이미지에 적용하는 실험!
- 이미지를 패치별로 쪼개고, 패치들의 linear embeddings sequence를 transformer의 input으로 넣음!
- 이미지 분류에 대한 모델을 supervised 방식으로 학습!

→ 중간 데이터셋에서는 낮은 정확도 (because of inductive biases의 부족)
But, 대용량의 데이터 셋에서는 good! 큰 스케일의 학습이 inductive bias를 이겨버린다!

### ▼ 3. 이 연구의 접근 방법은 무엇인가?(Method)

1. **Vision Transformer(ViT)**

Image를 patch로 쪼갠다는 개념에서 시작!

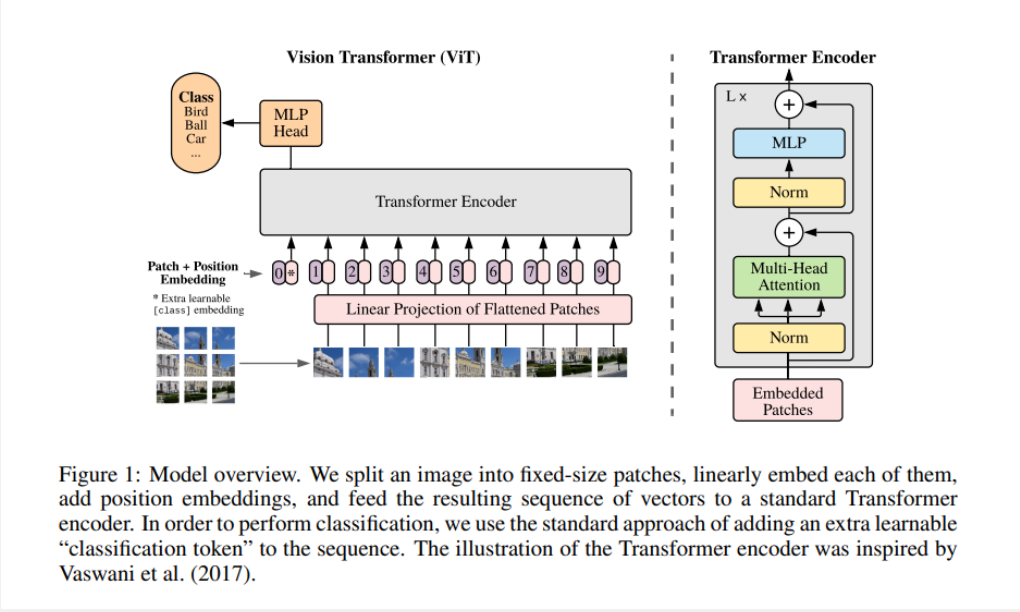



Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification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

1. 이미지를 16 x 16 patch의 크기로 분할
2. 16x16 patch로 잘라진 이미지들 각각 linear projection 수행
3. Positional encoding을 더해줌!
4. Transformer encoder를 통해 embedding 생성
5. [class]에 대한 embedding만 이용해 MLP에 넣어주고 classification 진행!

- **Inductive bias**
  **-** CNN이나 RNN의 경우 global 한 영역의 처리는 어렵
  - ViT는 일반적인 CNN과 다르게 공간에 대한 inductive bias가 없으므로, ViT는 더 많은 데이터를 통해 원초적인 관계를 robust 하게 학습시켜야 함!
- **Hybrid Architecture**
  - Image Patch의 대안으로, CNN을 통과한 feature map을 input sequence로 넣어줌!

2. **Fine-tuning and Higher resolution**

💡 - 사전 학습된 prediction head를 제거하고 0으로 초기화된 DxK feed-forward layer를 붙여줌!
- 사전학습 시보다 더 높은 해상도의 이미지로 fine-tuning하는 것이 ㄷ더 좋은 결과를 가져옴!
- 이러한 해상도 조정과 패치 추출이 ViT에서 수동적으로 image-specific inductive bias를 추가해주는 부분!

**▼ 4. 실험은 어떻게 이루어졌는가? (Experiments)**

💡 **1. SOTA와 비교!**
- 중간 사이즈 데이터 셋 : Baseline > ViT
- 큰 사이즈 데이터셋 : ViT > Baseline
- ViT의 경우 모델 파라미터가 증가하고, 데이터셋의 크기가 증가할 수록 지속적으로 성능이 증가함!

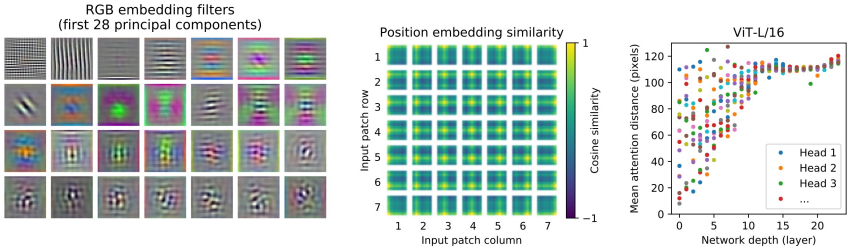
**2. Pre-training Data Requirements**
사전학습 시 사용되는 데이터셋 크기가 어떤 영향을 미치는지 실험
- ViT는 inductive bias가 기존 CNN보다 부족하므로, 사전학습 데이터셋의 크기가 클 때 좋은 성능 보여줌!
- 중간 데이터셋은 모델을 학습시키에 충분하지 않다!

**3. Scaling Study**
- ViT는 ResNet보다 동일한 성능을 내기 위해 반 정도의 컴퓨팅이 필요

-CNN의 feature map을 이용한 하이브리드 모델은 적은 computing cost에서는 ViT를 능가하지만 cost를 늘리게 되면 큰 차이 없음

-ViT는 saturate(포화)되지 않으면서 스케일링이 가능

**4. Inspecting Vision Transformer**
- Vision Transformer의 Embedding filter : 저차원의 CNN filter 기능과 유사
- Position Embedding 간의 유사성 : 가까운 패치 간의 유사도가 높다
→ Input Patch 간의 공간 정보가 잘 학습됨!
- Self Attention을 활용해 전체 이미지 정보의 통합 가능 여부 확인:




ViT는 가장 하위의 layer에서도 전체 이미지에 대한 정보를 통합할 수 있다! 낮은 layer의 self-attention head 는 CNN처럼 localization 효과를 보임!

**5. Self-Supervision**
- self-supervised 사전학습을 한 모델은 supervised 사전학습을 한 모델보다 4% 성능 하락

**▼ 5. 결론 및 요약 (Conclusion)**

💡 - 기존 연구와 달리 Image Recognition 분야에 트랜스포머를 직접적으로 적용한 사례
- 이미지에 특정된 inductive bias를 아키텍처에 주입하지 x
- 큰 사이즈의 데이터셋에서 사전 학습 한 후 기존의 CNN 기반 baseline보다 더 좋은 성능을 얻음!

* Challenge
1. Image recognition외에 detection, segmentation과 같은 task에 적용
2. Self - Supervision과 같은 사전 학습 방법에 대한 연구 필요
3. ViT가 포화되지 않았으므로 더 확장시켜 더 높은 성능을 향상할 필요