YONG IN UNIVERSITY

Attention & Transformer

[1706.03762v7] Attention Is All You Need

2017 NIPS

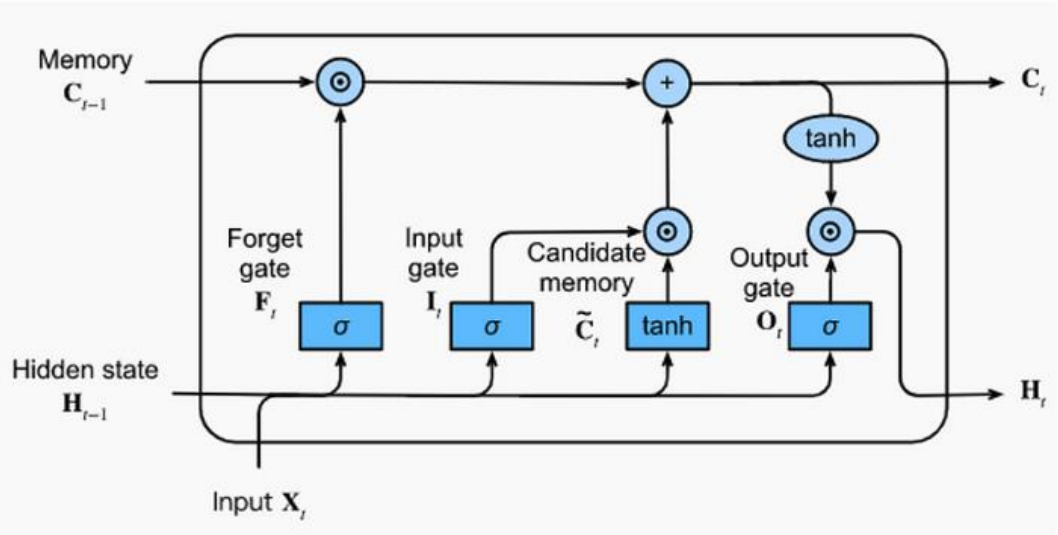
[2010.11929v2] AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

2021 ICLR

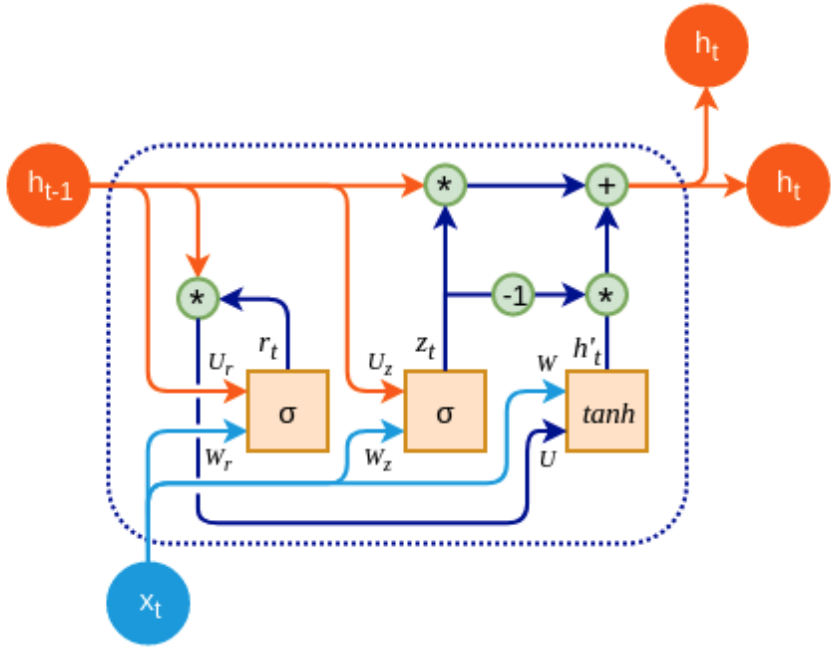
Presenter : Jeongbeen Son

1. Background & Motivation

YONG IN UNIVERSITY



LSTM

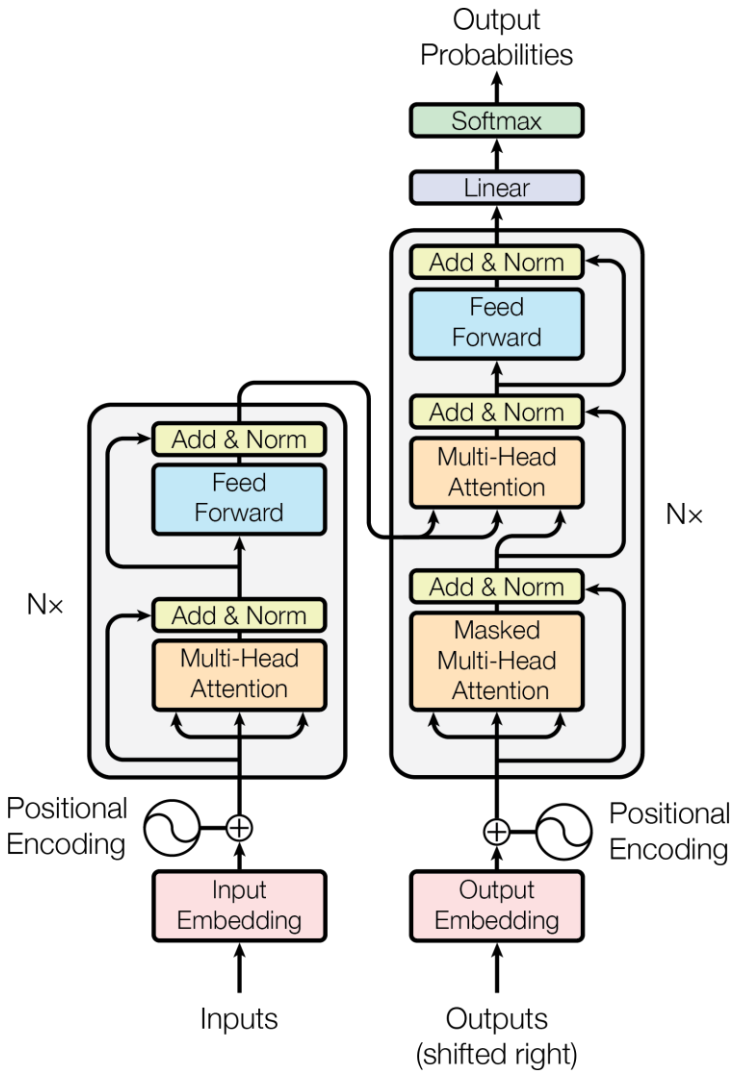


GRU

- 내재적인 계산 병목 현상
- 순차적 특성으로 인한 병렬화 불가

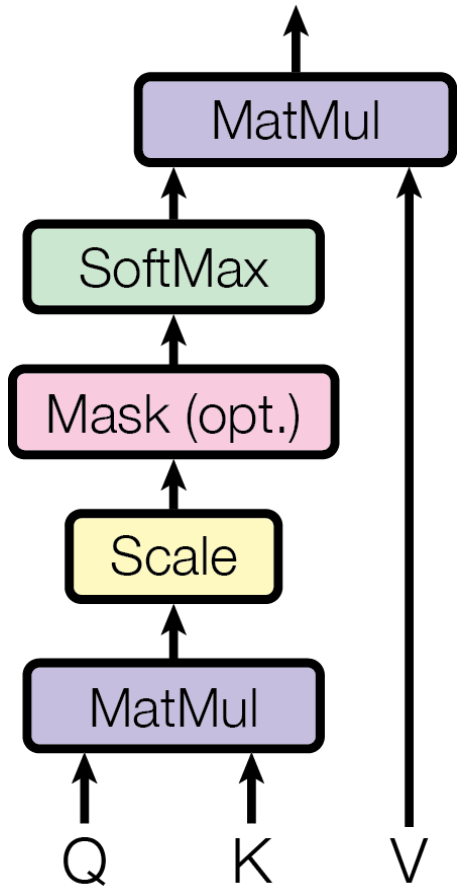
- Gradient Vanishing
- 장거리 의존성 학습의 어려움

2. Model Architecture



- Encoder-Decoder Framework
  - 6 layers in each Encoder and Decoder
    - MHSA(Encoder)
    - Masked MHSA + Encoder-Decoder Attention(Decoder)
    - Feed-forward
    - Residual connections
    - Layer normalization

3. Attention Mechanism



Self-Attention은 "모든 위치 쌍 (i, j)의 관계를 계산해,
i의 표현을 j들의 가중합으로 갱신하는" **전역(pairwise)** 상호작용 연산이다.

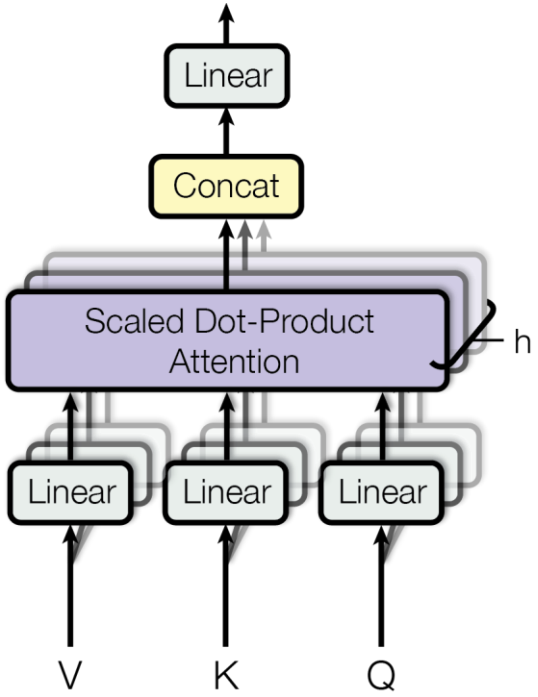
- Base operation = "weighted sum of values"

  각 토큰/패치 i는 다른 모든 토큰/패치 j를 **참조**해서 출력을 만든다.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

  여기서 softmax가 "i가 j를 얼마나 볼지 정하고, 그 비율로 V를 섞는다.

3. Attention Mechanism



- Q/K/V는 "입력의 서로 다른 역할" 로 선영 투영된 것

  입력 특징 x에서 Query/Key/Value를 학습된 선형 변환으로 만든다(투영 행렬 W^Q, W^K, W^V).

  MultiHead는 이 투영을 여러 세트로 병렬 수행해, 서로 다른 표현 하위공간에서 동시에 관계를 학습한다.

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)W^O$$

$$\mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

  한 개의 head의 평균이 주는 표현 손실을 완화하는 설계

4. Positional Encoding

Transformer는 recurrence/convolution이 없기 때문에 sequence에 대한 이해가 없음.
이를 해결하기 위해 모델은 입력 임베딩에 추가되는 위치 인코딩을 통합함.

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{model}}\right)$$

$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d_{model}}\right)$$

Transformer의 Positional Encoding은 삼각함수의 덧셈 정리(Trigonometric Addition Theorems)를 이용하여 고정된 오프셋 k에 대하여 PE_{pos+k}를 PE_{pos}의 선형 함수로 표현함.

4. Positional Encoding

$$\begin{pmatrix} PE_{(pos+k,2i)} \\ PE_{(pos+k,2i+1)} \end{pmatrix} = \begin{pmatrix} \cos(\omega_i k) & \sin(\omega_i k) \\ -\sin(\omega_i k) & \cos(\omega_i k) \end{pmatrix} \begin{pmatrix} PE_{(pos,2i)} \\ PE_{(pos,2i+1)} \end{pmatrix}$$

위치 벡터 $(PE_{(pos,2i)}, PE_{(pos,2i+1)}$ 에 특정 각도 $\omega \cdot k$만큼 회전시키는 행렬을 곱하면

k 만큼 떨어진 위치의 벡터를 얻음

이 성질을 통하여 신경망은 절대적인 위치 값 pos를 직접 알지 못하더라도,

가중치 행렬을 통해 이러한 회전 변환을 학습함으로써

단어들 사이의 상대적인 거리 k 를 파악하고 활용함

5. Applications of Attention

- Encoder-Decoder Attention(Cross-Attention)

  이전 decoder layer → Query

  Encoder의 출력 → Key/Value

- Self-Attention

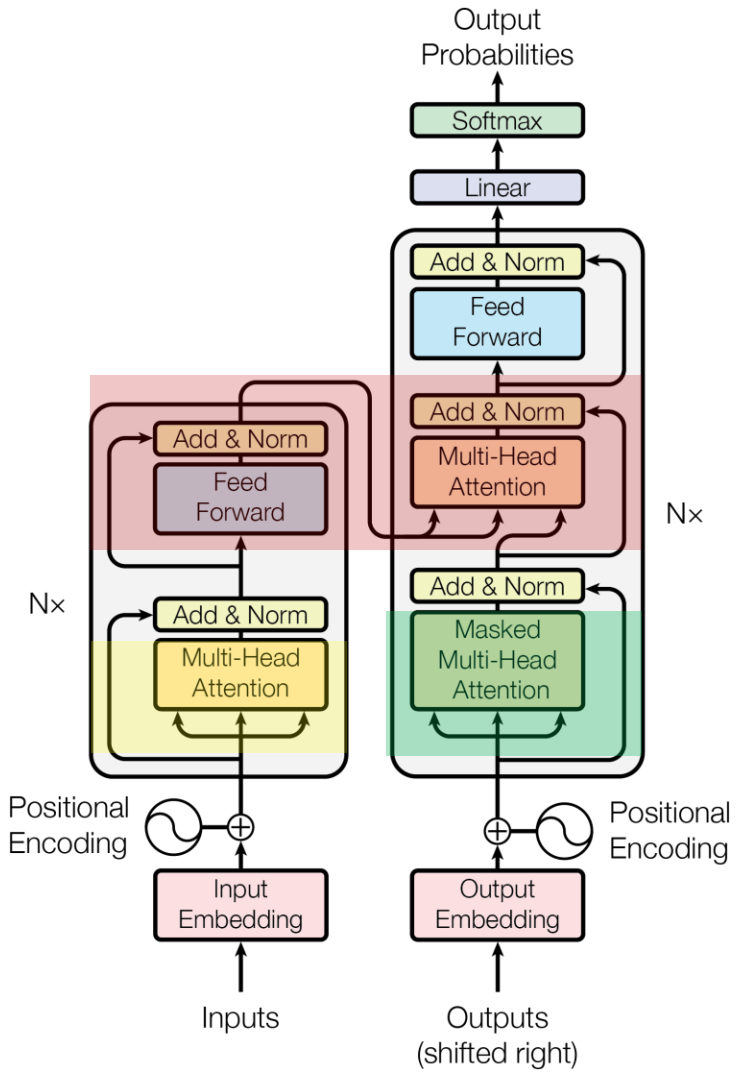
  Encoder

  이전 encoder layer → Query/Key/Value

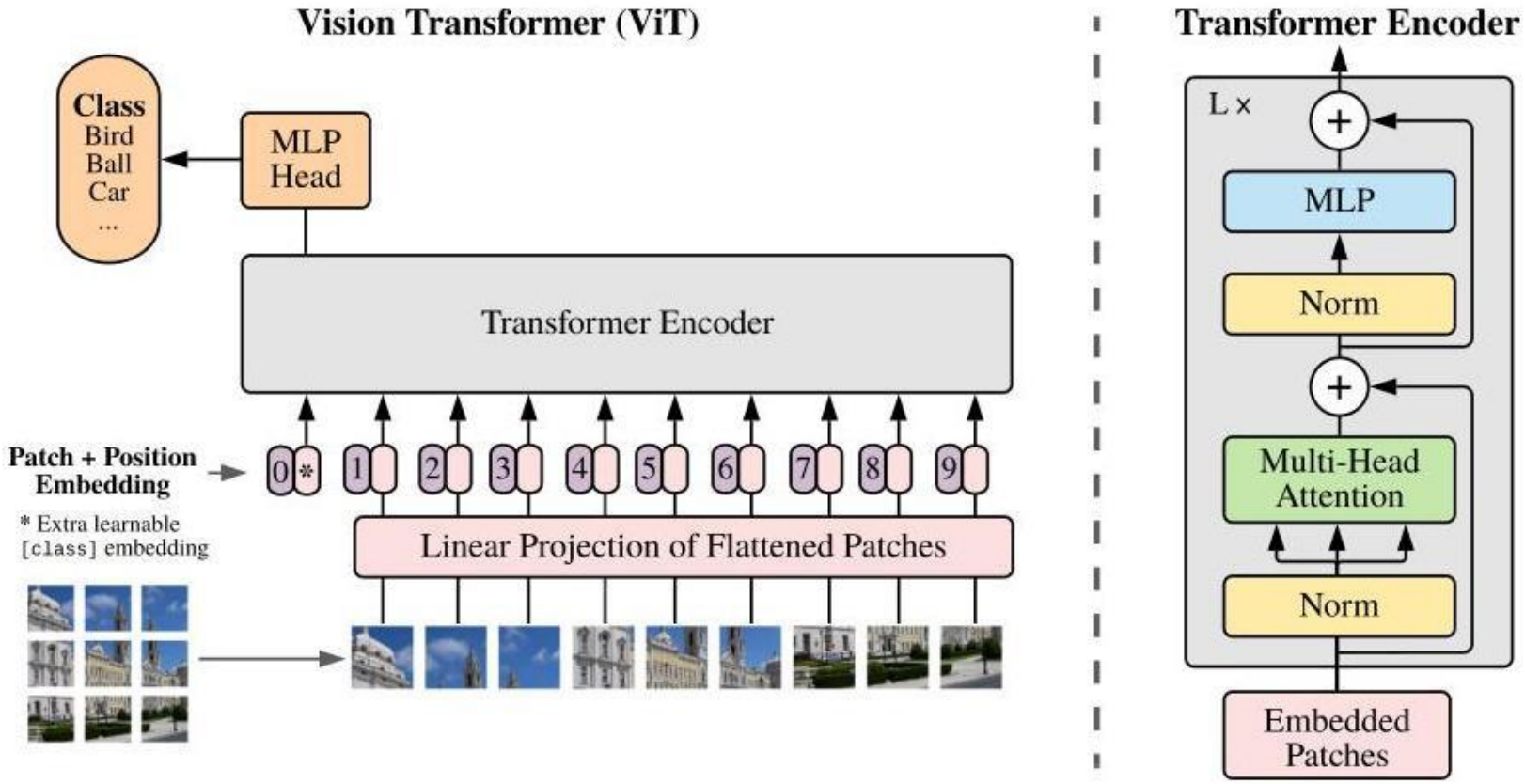
  Decoder

  이전 decoder layer → Query/Key/Value

  auto-regressive 속성으로 인한 정보의 흐름을 Masking으로 제한

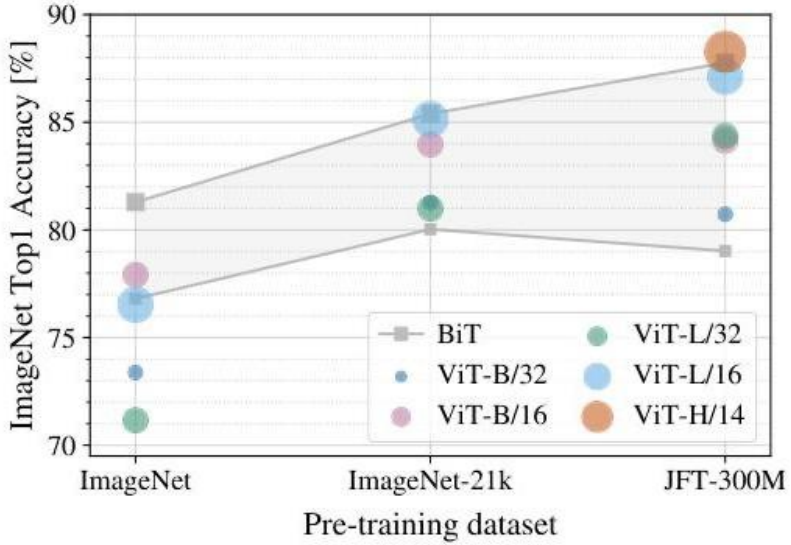


7. Vision Transformer



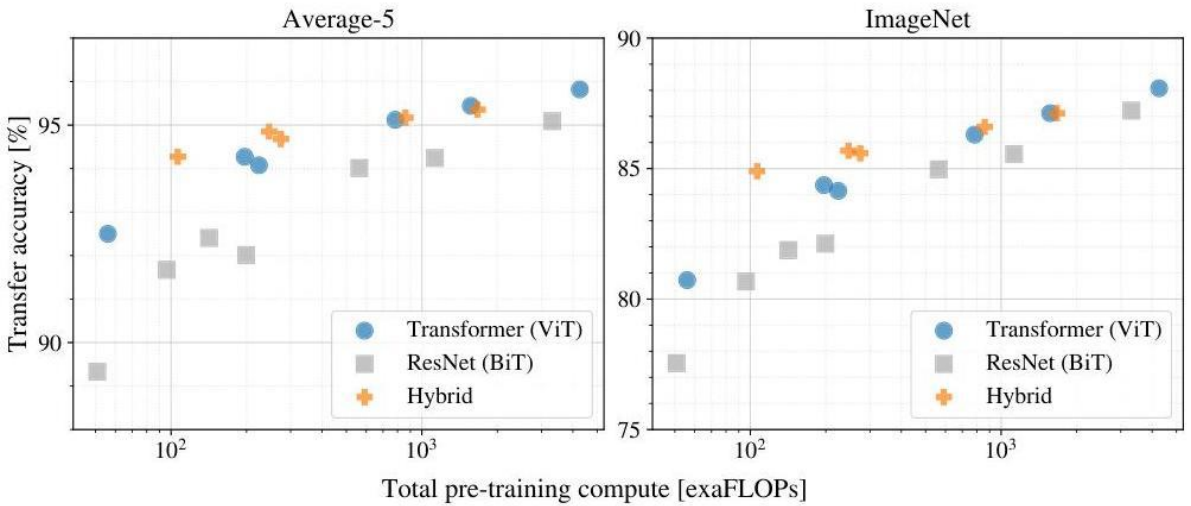
이미지 패치 처리와 시퀀스 구성을 통하여 Computer Vision Tasks에 Transformer를 적용함

8. Experiments



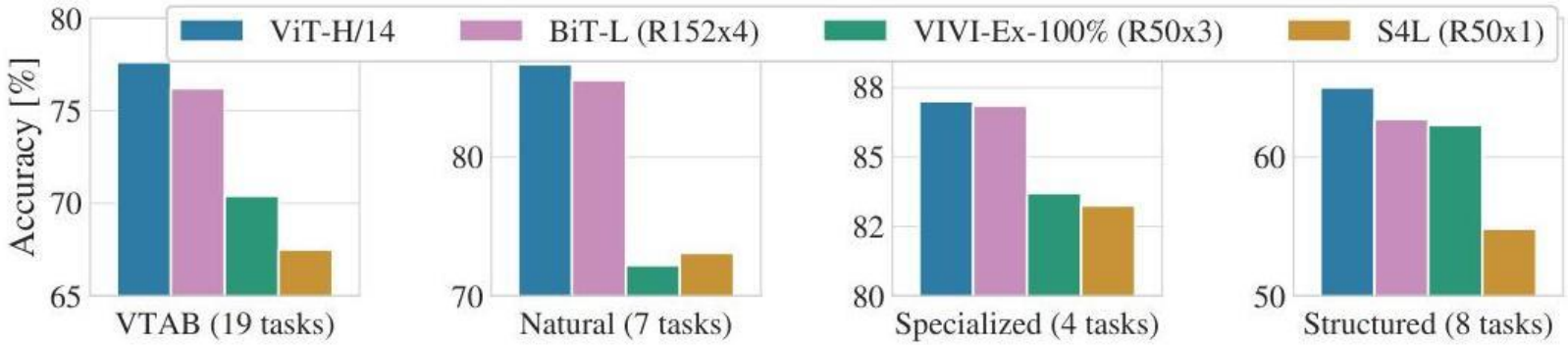
- 규모 의존적 성능
  데이터셋의 크기가 커질 수록 ResNet에 비해 ViT의 성능이 상승한다.

8. Experiments



- 계산 효율성
  더 나은 성능을 달성했음에도 ViT는 더 적은 계산 자원을 필요로 한다.

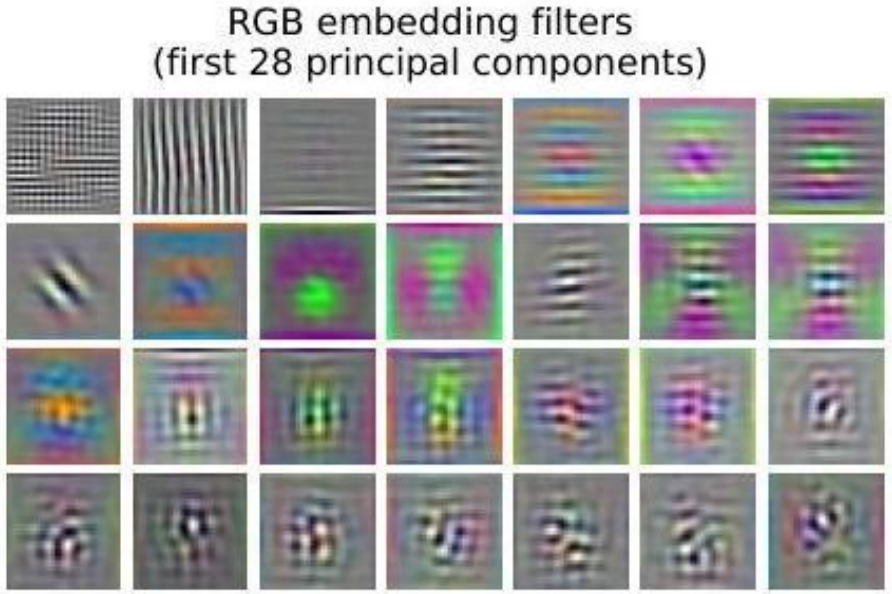
8. Experiments

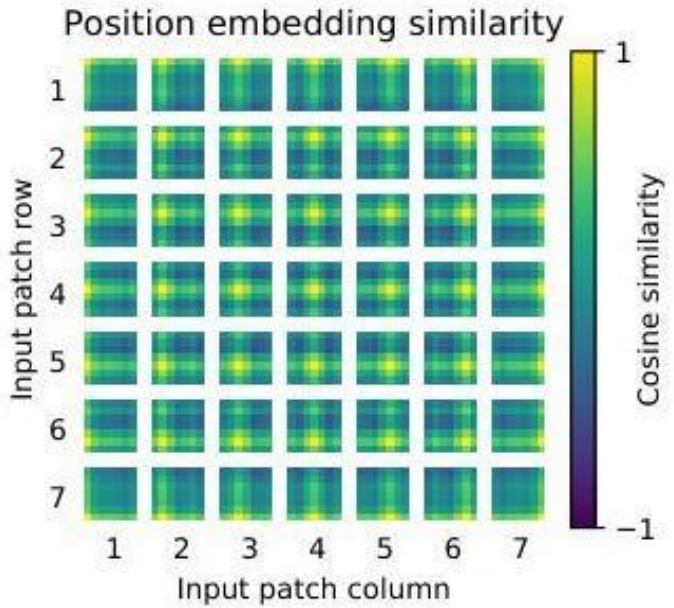


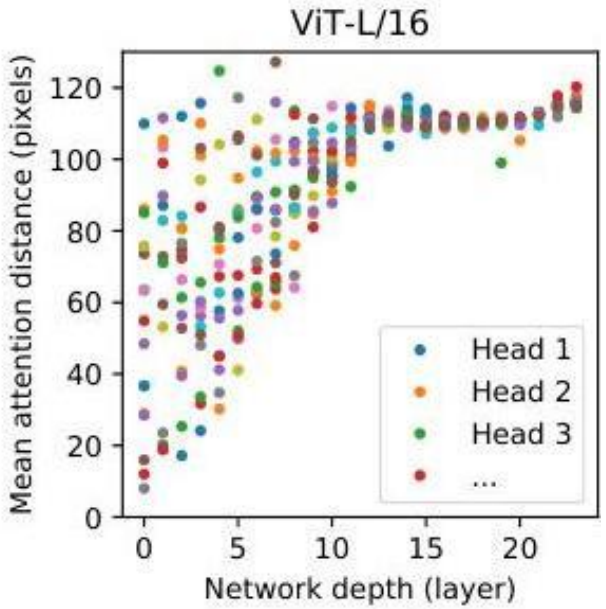
- 벤치마크 성능
  여러 평가 벤치마크에서 JFT-300M으로 사전 훈련된 모델은 이전의 모델들을 능가함

9. Interpretability and Internal Representations

YONG IN UNIVERSITY





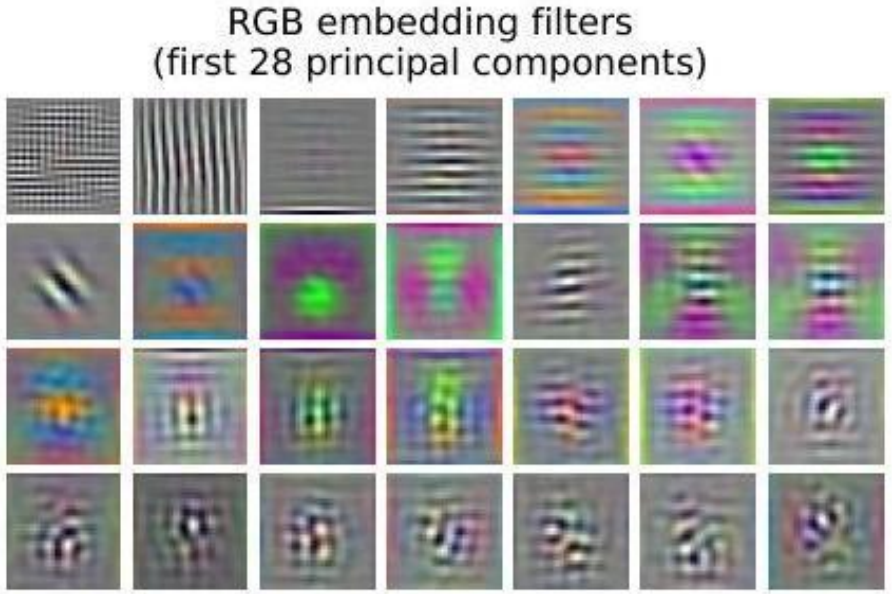


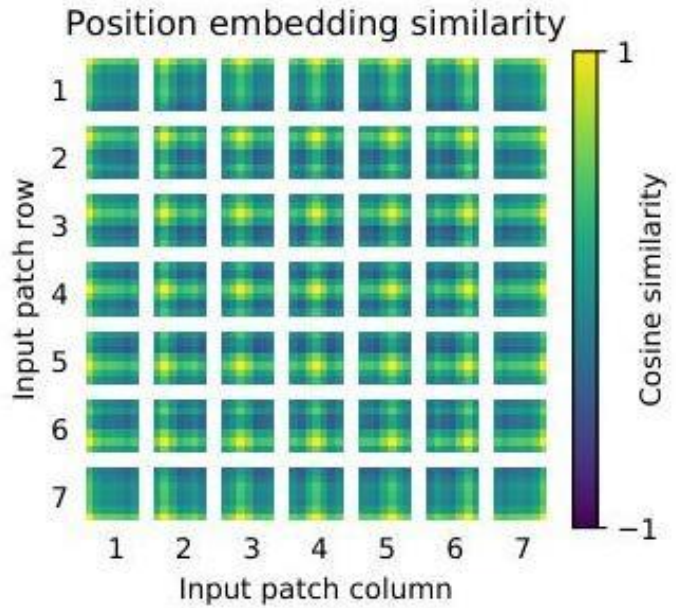
- Left

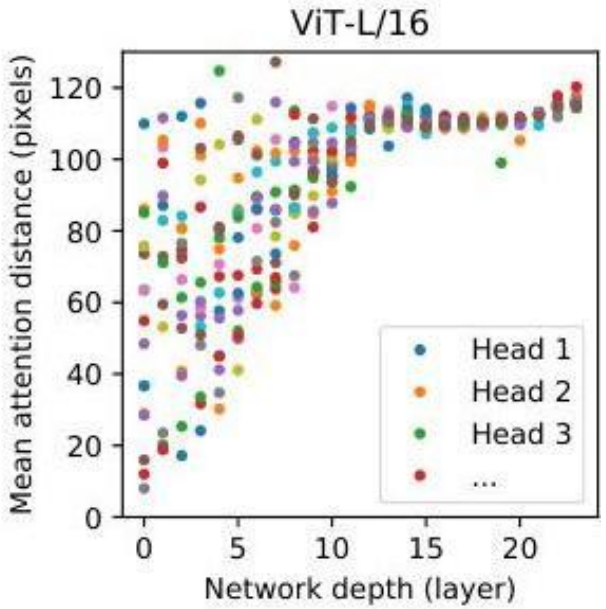
  ViT는 이미지를 Flatten하게 펴서 입력하지만
  CNN의 초기 레이어가 선과 색상 같은 저수준 특징을 학습하는 것과 유사하게
  첫 번째 필터에서 색상, 엣지 등을 표현하는 기저 함수 형태를 띄고 있음

9. Interpretability and Internal Representations

YONG IN UNIVERSITY



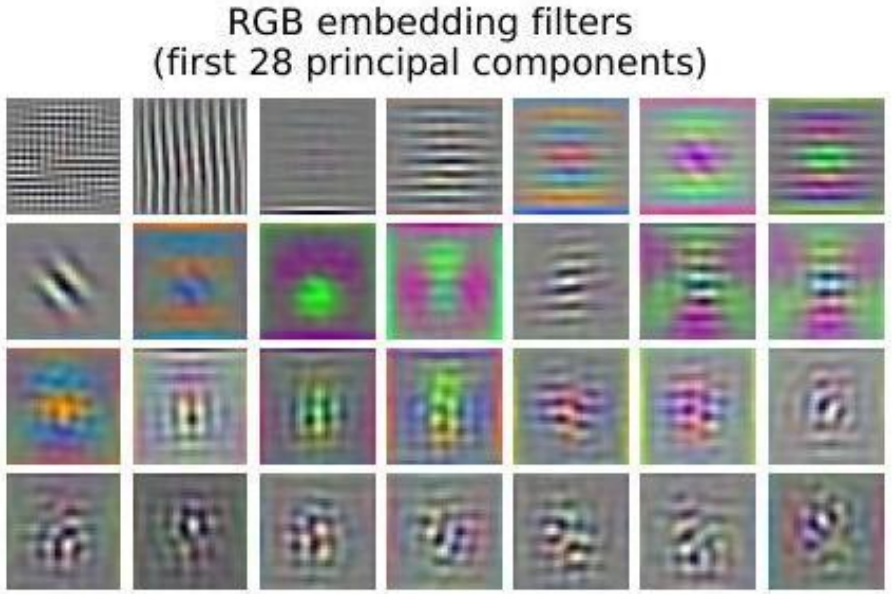


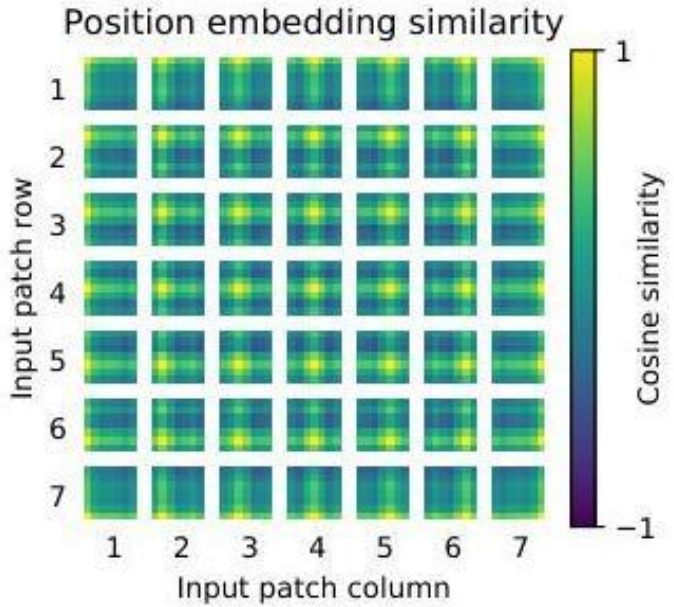


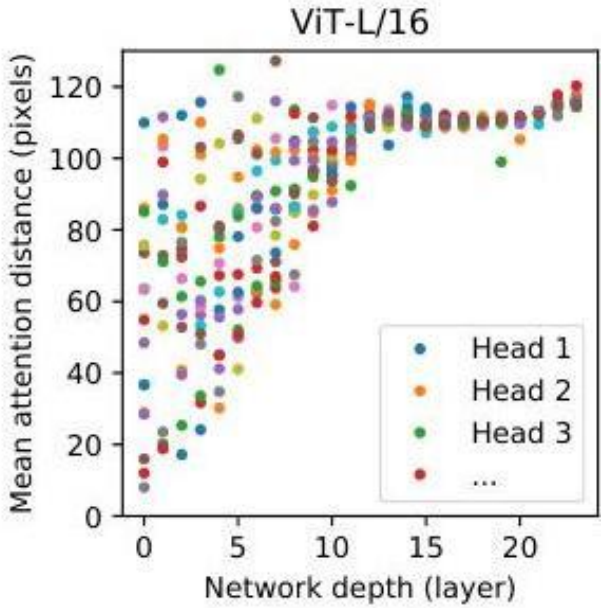
- Center

  가까운 패치일수록, 같은 행, 열에 있을수록 더 유사한 위치 임베딩을 갖는 경향이 있음
  모델이 스스로 이미지의 2D 구조를 파악하고 했음을 증명함
  모델이 위치 임베딩 간의 유사도를 통해 이미지의 거리 개념을 인코딩하도록 학습됨

9. Interpretability and Internal Representations







- Right

  하위 레이어의 어떤 헤드들은 CNN처럼 좁은 영역만 보고, 어떤 헤드들은 이미 전역을 봄
  레이어가 깊어질수록 전체를 보는 경향이 커짐

  ViT가 CNN의 구조적 제약 없이도, 스스로 이미지의 공간적 구조를 학습하고,
  국소적 정보와 전역적 정보를 동시에 처리할 수 있음

10. Impact and Significance

CNN이 고성능 이미지 인식에 필수적이지 않음을 입증함으로써 패러다임 전환을 나타냄

- Architectural Unification

  단일 아키텍처가 다양한 데이터 양식에 걸쳐 뛰어날 수 있음을 시사하며,
  잠재적으로 multi-modal 시스템의 개발을 단순화함

- Scale as Design Principle

  아키텍처 선택 시 훈련 데이터 및 계산 자원의 가용성을 고려해야하며,
  일반적인 아키텍처가 더 큰 규모에서 매력적이라는 점을 강조함

- Research Direction

  Object Detection, Segmentation, Video Understanding에 대한 응용뿐만 아니라
  ViT를 위한 자기 지도 학습 연구를 포함하여 Computer Vision 분야에서 새로운 연구방향을 제시함