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

https://arxiv.org/pdf/2010.11929.pdf

Abstract

Transformer 구조가 자연어 처리 작업에서는 표준이 되었지만, 컴퓨터 비전 분야에서의 적용은 제한적이다.

하지만 본 논문에서는, CNN에 의존이 필수적이지 않고 오로지 Transformer만으로도 이미지 분류 작업에서 매우 우수한 성능을 낼 수 있었음을 확인하였다.

Introduction

자연어 처리(NLP)에서는 self-attention 기반 구조, 특히 Transformer가 선호되는 모델로 자리 잡았다. 이 모델은 대규모 text corpus에서 사전 학습을 진행한 후, 더 작은 dataset에서 fine-tune하는 방식으로 사용된다. Transformer의 계산 효율성과 확장성 덕분에, 100B가 넘는 매개변수를 가진 모델을 훈련시킬 수 있게 되었으며, 모델과 dataset이 커지더라도 성능이 감소하지 않았다.

그러나 컴퓨터 비전 분야에서는 여전히 CNN 구조가 우세하다. NLP의 성공으로 인해 self-attention과 CNN의 결합을 시도했으며, 일부는 CNN을 완전히 대체하기도 했다. 그러나 이런 모델들은 이론적으로 효율적이지만, attention 패턴 사용으로 인해 현대 하드웨어

가속기에서 효과적으로 최적화되지 못했다. 따라서 large-scale 이미지 인식에서는 전통적인 ResNet 같은 구조가 여전히 최고의 기술로 남아 있다.

본 논문에서는 이를 개선하고자, 가능한 적은 수정으로 이미지에 직접 Transformer를 적용하는 실험을 진행했다. 이를 위해 이미지를 patch로 나누고, 이 patch들의 linear embeddings sequence를 Transformer에 입력으로 제공한다. 이미지 patch는 NLP 에서 토큰(단어)처럼 취급된다. 이 모델은 지도 학습 방식으로 이미지 분류 작업에 훈련된다.

ImageNet과 같은 중간 크기의 데이터셋에서 강력한 정규화 없이 훈련했을 때, 이 모델들은 비슷한 크기의 ResNet보다 몇 퍼센트 포인트 낮은 정확도를 보인다. → Inductive bias

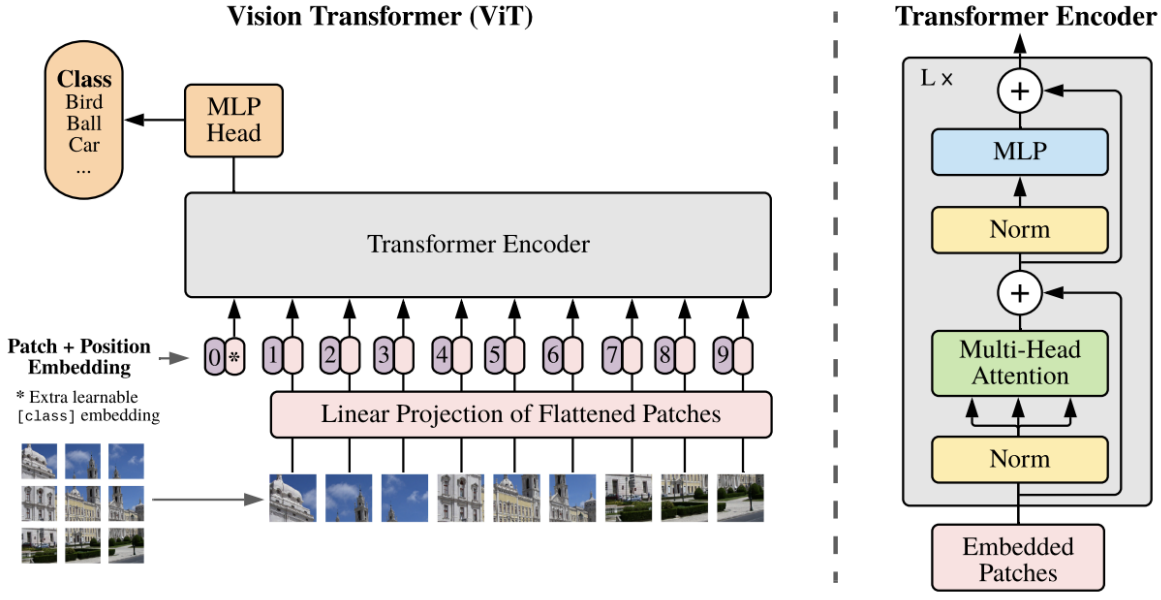
Inductive bias

: 본 적 없는 입력의 출력을 잘 예측하기 위해 사용하는 가정

Transformer는 Inductive bias가 부족하다. 이것은 translation equivariance와 locality 같은 것인데 그렇기에 데이터가 불충분할 때 generalization 하기가 힘들다고 설명한다. 다만 large-scale 데이터로 했을 때는 이것을 이겨낼 수 있었다.

Method

Vision Transformer (ViT)



원래의 Transformer 모델은 1차원 토큰 임베딩의 시퀀스를 입력으로 받는데, 2차원 이미지를 다루기 위해, 이미지를 여러 개의 2차원 패치로 재구성한다. 여기서 이미지 x는 H×W×C 차원을 가지며, H와 W는 각각 이미지의 높이와 너비, C는 채널의 수를 의미한다. 각 이미지 패치는 P×P의 해상도를 가지며, 이 패치들을 평탄화하여 N×(P^2·C) 차원의 시퀀스로 변환한다. 여기서 N은 패치의 총 개수로, Transformer에 입력되는 효과적인 시퀀스 길이가 된다.

Transformer는 모든 레이어에서 일정한 잠재 벡터 크기 D를 사용하므로, 패치들을 평탄화하고 학습 가능한 선형 투영을 통해 D 차원으로 매핑한다. 이 투영된 패치들을 패치 임베딩이라고 한다.

BERT의 [class] 토큰과 비슷하게, 임베딩된 패치 시퀀스 앞에 학습 가능한 임베딩을 추가한다. 이 임베딩의 상태는 Transformer 인코더의 출력에서 이미지 표현으로 사용된다. 사전 학습과 미세 조정(fine-tuning) 모두에서, 분류를 위한 헤드가 이 임베딩에 연결된다. 사전 학습 시에는 하나의 은닉층을 가진 MLP로, 미세 조정 시에는 단일 선형 층으로 구현된다.

패치 임베딩에는 위치 정보를 유지하기 위해 위치 임베딩이 추가된다. 여기서는 더 복잡한 2차원 위치 임베딩보다는 표준 학습 가능한 1차원 위치 임베딩을 사용하는데, 이는 더 복잡한 방식에서 유의미한 성능 향상을 관찰하지 못했기 때문이다.

Transformer 인코더는 multiheaded self-attention(MSA)과 MLP 블록의 교대 레이어로 구성되며, 각 블록 전에는 Layernorm(LN)이 적용되고, 각 블록 후에는 잔차 연결이 적용된다. 또한 actication function은 GELU 함수를 사용했다. 아래는 전체적인 과정을 수식으로 나타낸 것이다.

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}};\ \mathbf{x}_p^1\mathbf{E};\ \mathbf{x}_p^2\mathbf{E};\ \cdots\ ;\ \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, \qquad \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D},\ \mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D} \tag{1}$$

$$\mathbf{z}'_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \qquad \ell = 1 \ldots L \tag{2}$$

$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \qquad \ell = 1 \ldots L \tag{3}$$

$$\mathbf{y} = \text{LN}(\mathbf{z}_L^0) \tag{4}$$

Inductive bias

ViT는 CNN에 비해 이미지에 특화된 inductive bias가 적다. CNN에서는 locality, 2D neighborhood structure, 그리고 translational equivariance가 모델 전체의 각 레이어에 내장되어 있다. 반면, ViT에서는 MLP 레이어만이 local하고 translational equivariant하며, self-attention 레이어는 global하다. 2D neighborhood structure는 매우 제한적으로 사용되는데, 모델의 시작 부분에서 이미지를 패치로 나누는 과정(patch extraction)과, 다른 해상도의 이미지에 대한 위치 임베딩을 조정(resolution adjustment)하는 fine-tuning 시간에만 사용된다. 그 외에는 초기화 시점의 위치 임베딩이 패치의 2D 위치에 대한 정보를 전혀 가지고 있지 않으며, 패치 간의 모든 공간적 관계는 처음부터 학습되어야 한다.

Hybrid Architecture

원본 이미지 패치 대신, CNN에서 추출한 feature map으로부터 입력 시퀀스를 형성할 수 있다. 이 하이브리드 모델에서는 패치 임베딩 투영 E가 feature map에서 추출된 패치에 적용된다. 특별한 경우로, 패치는 공간 크기가 1x1일 수 있는데, 이는 입력 시퀀스가 단순히 feature map의 공간 차원을 평탄화하고 Transformer 차원으로 투영함으로써 얻어진다는 의미이다.

Fine-tuning and Higher Resolution

일반적으로 ViT는 큰 데이터셋에서 pre-train을 하고, 다음에 downstream task에 fine tuning을 한다. 이를 위해, pre-train한 예측 head를 제거하고 0으로 초기화된 D × K feedforward network를 연결한다(K는 하류 클래스의 수).

pre-train보다 높은 해상도에서 fine-tuning하는 것이 종종 유익하다. 더 높은 해상도의 이미지를 입력할 때, 패치 크기는 동일하게 유지되어 더 큰 효과적인 시퀀스 길이를 결과로 낳는다. Vision Transformer는 임의의 시퀀스 길이를 처리할 수 있지만(메모리 제약에 따라), 사전 학습된 위치 임베딩은 더 이상 의미가 없을 수 있다. 따라서 원본 이미지에서의 위치에 따라 사전 학습된 위치 임베딩의 2D 보간을 수행한다.

Conclusion

Transformer 기술을 이미지 인식에 직접 적용해보면서, 이미지를 작은 패치들의 연속으로 보고 이를 표준 Transformer 인코더로 처리하는 새로운 방식을 탐구했다. 이 방법은 큰 데이터셋에서 사전 학습할 때 매우 효과적이며, 많은 이미지 분류 작업에서 최고의 성능을 달성하거나 넘어서면서도 비용 효율적이다. 그러나 ViT를 다른 컴퓨터 비전 작업에 적용하거나, self-supervised 사전 학습 방법을 더 발전시키고, ViT를 더 확장하는 등 아직 극복해야 할 도전 과제들이 남아 있다.