ViT

김동원

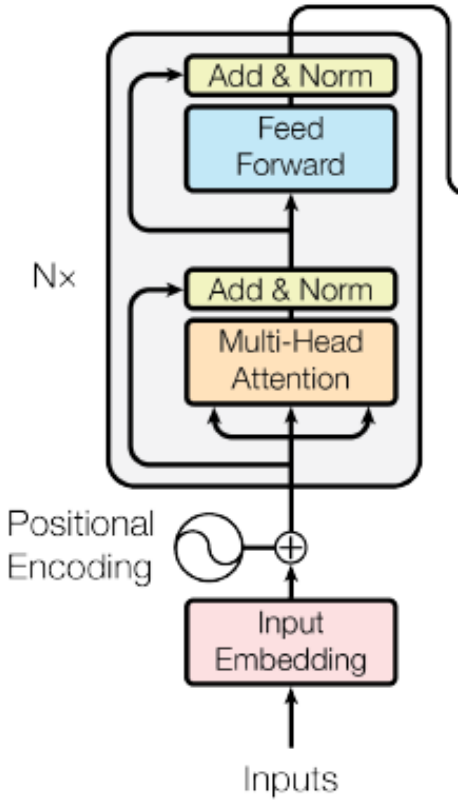
목차

ViT

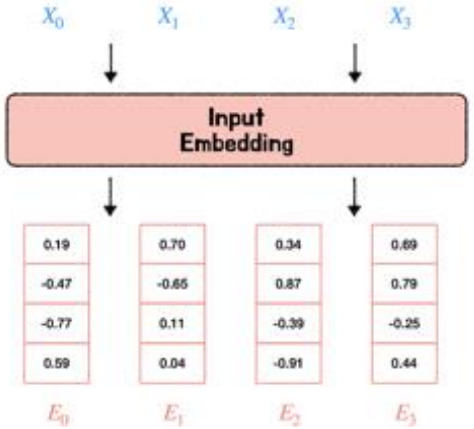
ViT(An image is worth 16x16 words: transformer for image recognition at scale)

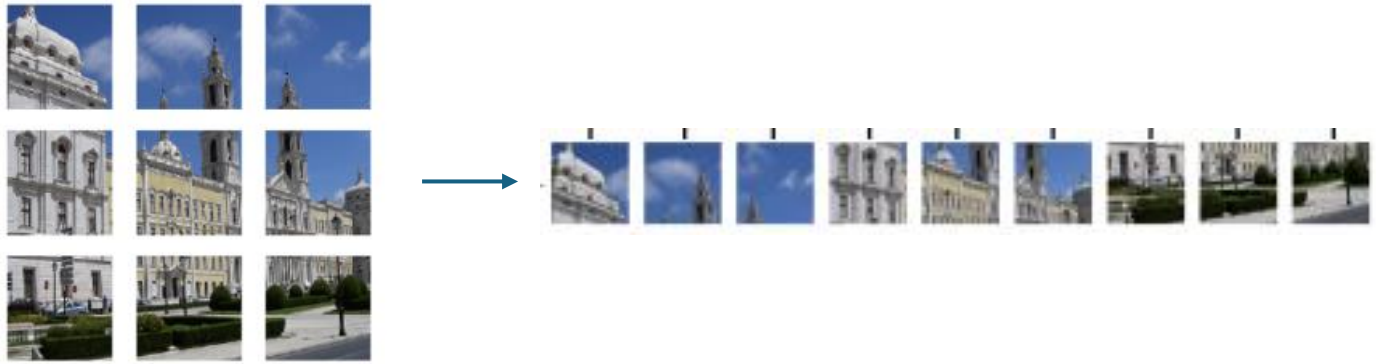
방법

이미지를 작은 조각으로 나누어 pixel들을 일자로 나열하여 transformer는 토큰처럼 입력을 받는다.



Transformer

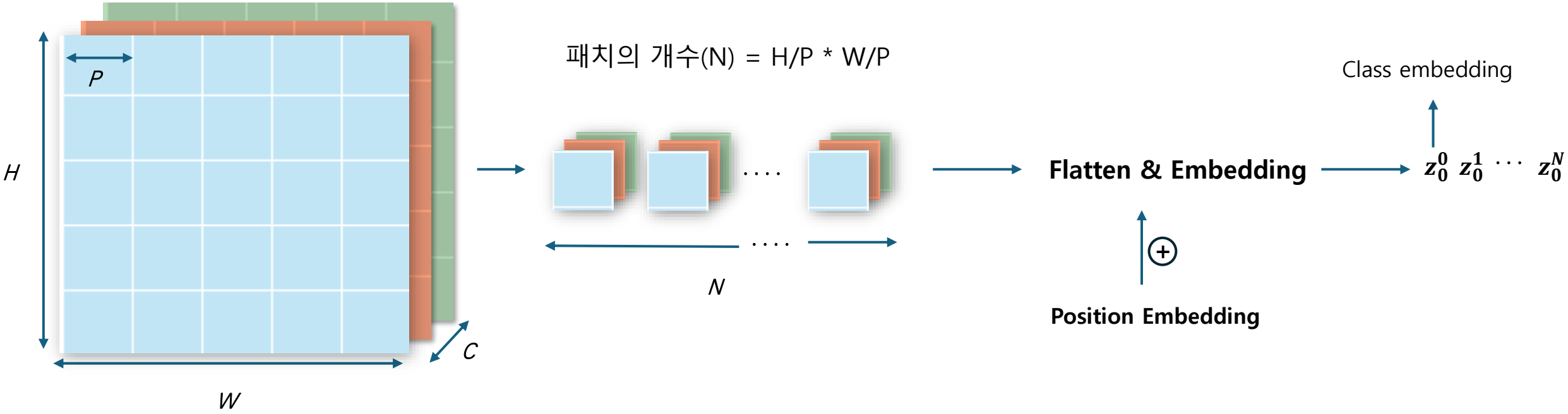




ViT(An image is worth 16x16 words: transformer for image recognition at scale)

방법

이미지를 작은 조각으로 나누어 pixel들을 일자로 나열하여 transformer는 토큰처럼 입력을 받는다.



ViT(An image is worth 16x16 words: transformer for image recognition at scale)

Positional embedding

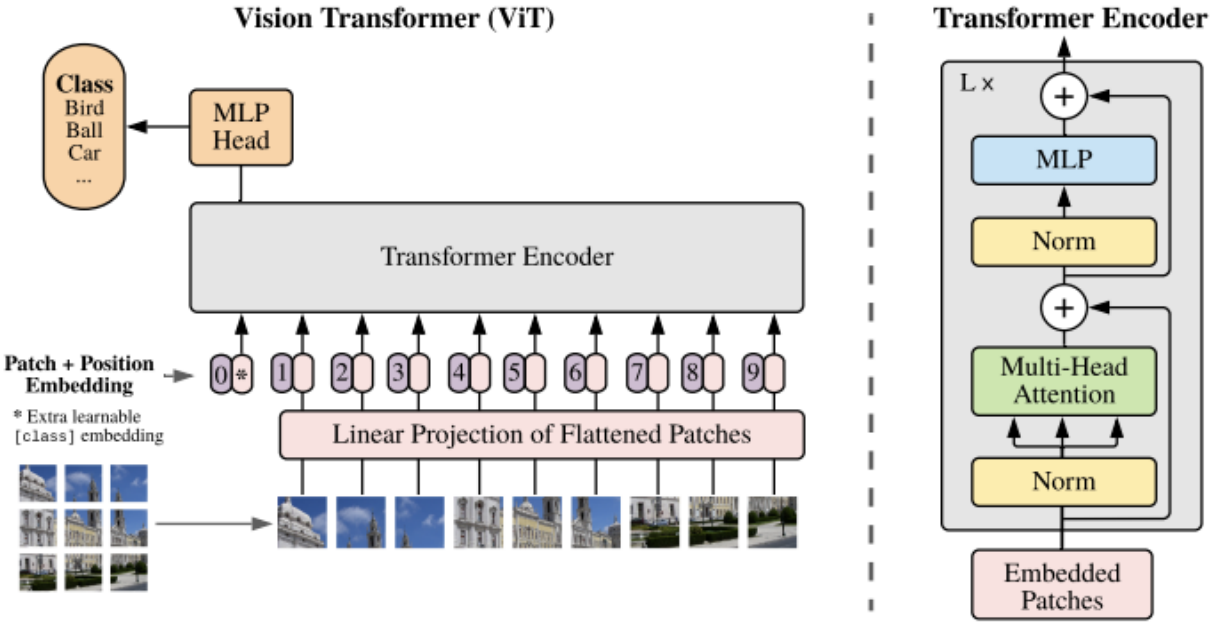
Positional embedding의 방식은 여러 방법을 테스트하고 성능이 좋은 방법을 택하였다.

1. No positional information
2. 1D positional embedding: 이미지 조각을 수평적으로 나열한 순서를 사용
3. 2D positional embedding: 이미지의 조각을 수직, 수평 위치를 각각 절반만큼 embedding하여 학습하는 방법
4. Relative positional embedding: 절대적 위치가 아닌 이미지 조각 간의 상대적 거리를 사용

Pos. Emb.	Default/Stem	Every Layer	Every Layer-Shared
No Pos. Emb.	0.61382	N/A	N/A
1-D Pos. Emb.	0.64206	0.63964	0.64292
2-D Pos. Emb.	0.64001	0.64046	0.64022
Rel. Pos. Emb.	0.64032	N/A	N/A

Positional embedding을 사용 여부는 중요하지만 어떤것을 사용하는 지는 중요하지 않았다. 이것은 입력을 이미지의 조각 단위로 받아서 픽셀로 입력을 받았을 경우보다 위치정보가 중요하지 않다고 추측함.

ViT(An image is worth 16x16 words: transformer for image recognition at scale)



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

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

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

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

· Transformer encoder는 class embedding의 결과를 가지고 class를 분류한다.

·ViT는 대규보의 사전학습과 적은 데이터로 진행하는 Fine tuning 단계를 거친다.

·ViT는 대규모의 사전학습을 하고, 더 높은 해상도를 가진 작은 데이터로 fine tuning을 수행하고, 이때 patch 크기는 동일하게 유지하므로 입력의 길이는 길어지고, 이에 따라 positional embedding에서 2D interpolatio을 거친 것을 사용한다.

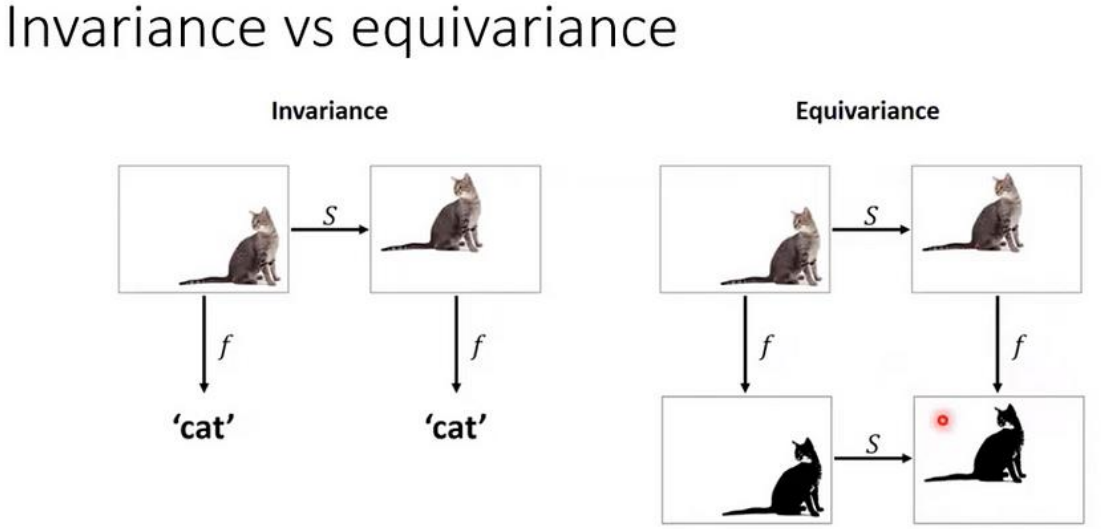
ViT(An image is worth 16x16 words: transformer for image recognition at scale)

Inductive Bias & Hybrid Architecture

·ViT는 inductive bias가 부족하여 학습량이 적은 경우 성능이 하락하지만 데이터를 충분히 주어졌을 때에는 좋은 성능을 보여줬다.

·Inductive bias란 귀납적 편향을 의미하며, 과제를 수행하기 위해 추가적인 가정을 의미,
CNN 구조에서는 이미지에서 주변의 픽셀들끼리 영향이 있다는 가정을 구조적으로 내포한다

·ViT에서는 inductive bias를 위해 position embedding과 이미지를 작은 조각으로 나누어 입력하였다. 하지만 position embedding은 학습을 통해 얻게 되므로 초기에는 정보가 부족하고, 사전학습과 fine tunning과정의 데이터의 해상도의 차이로 position embedding은 아주 제한적으로 사용된다.

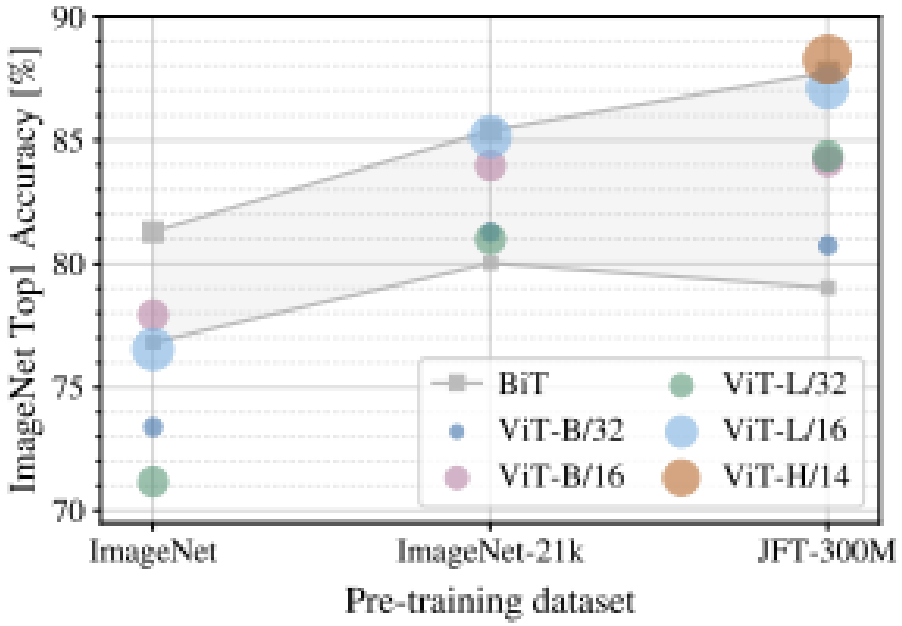


·이의 대안으로 hybrid architecture이 제안되었다. 이는 이미지를 CNN에 적용시키고 그 결과를 패치로 나누어 ViT의 입력으로 갖는 것이다.

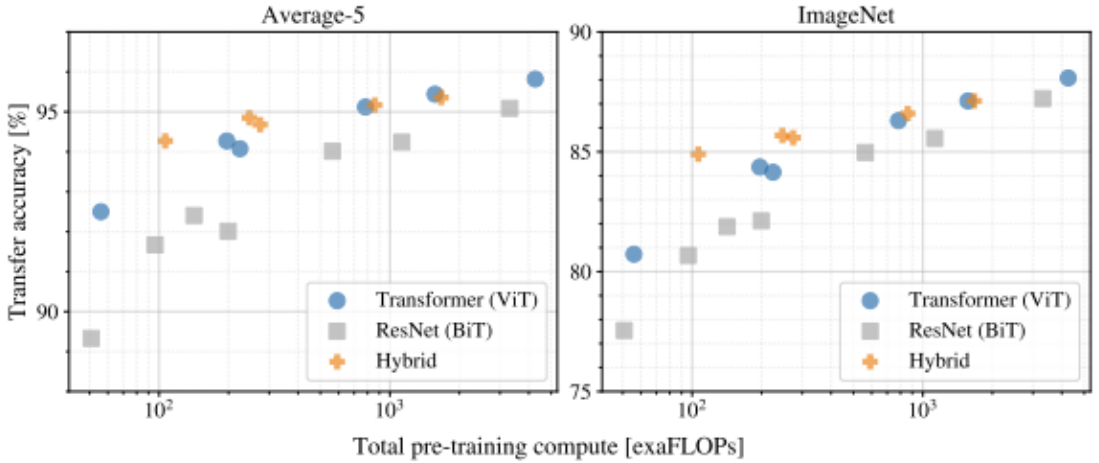
·실험을 통해 확인한 결과 inductive bias는 학습량을 늘리는 방법으로 극복할 수 있고 오히려 inductive bias가 적은 pure ViT 모델이 좋은 성능을 보여줌

ViT(An image is worth 16x16 words: transformer for image recognition at scale)

실험



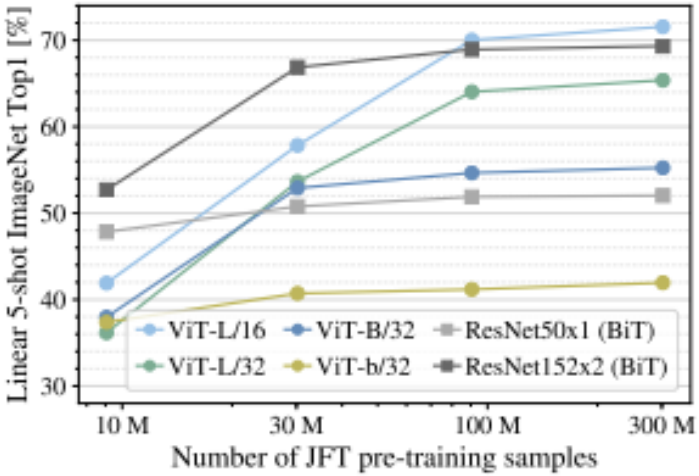
·큰 모델을 사용하는 것은 큰 데이터에서 학습시킬 경우에만 효과가 효과적이다.



·CNN모델 보다 동일한 연산량에서 성능이 좋다.

ViT(An image is worth 16x16 words: transformer for image recognition at scale)

실험



·CNN 모델은 적은 데이터에서 좋은 성능을 보이지만 데이터가 커지면 빠르게 성능향상을 멈춤

·ViT모델은 적은 데이터에서는 성능이 나쁘지만 많은 데이터에서 학습하면 높은 성능을 보여줌

결과

·CNN 모델보다 연산효율이 좋다

·inductive bias가 부족하지만 큰 데이터에서 학습하는 경우 극복된다

·큰 모델은 데이터셋이 많은 경우 사용하는 것이 좋다.

참고자료

Translation invariant & equivariance 이미지