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

AI 키워드	
날짜	@2024년 10월 14일
콘텐츠	
태그	

ABSTRACT

cv에서 attention은 cnn과 결합해서 사용하거나 특정 부분만을 대체함

cnn없이 트랜스포머만 가지고 이미지(sequences of image patches) 분류 태스크 잘 수행

사전학습 → 전이 학습(ImageNet, CIFAR-100, VTAB, etc.) 시 sota 달성

학습 시 필요한 계산량 크게 줄어듦

METHOD

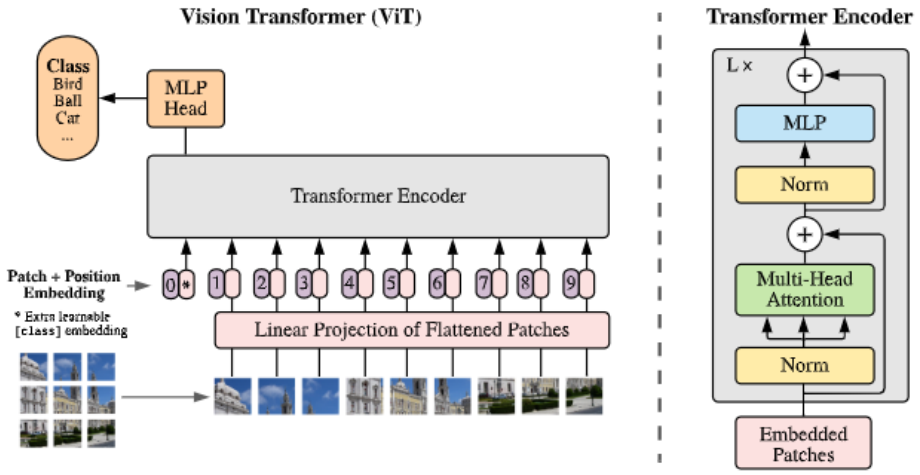


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification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

기존 트랜스포머 구조를 최대한 따름

VISION TRANSFORMER (VIT)

입력 embedding

1D 토큰 임베딩 시퀀스를 입력받으므로 2D 이미지를 flatten된 2D 패치들을 reshape 해야 함

$x \in R^{H \times W \times C} \rightarrow x_p \in R^{N \times (P^2 \cdot C)}$

H: height, W: width, C: channel, P: patch size

$N = \frac{HW}{P^2}$: 패치 개수 (=토큰 임베딩 시퀀스 길이)

flatten한 패치를 학습 가능한 linear projection(E)를 사용하여 D 차원으로 매핑

D (=d_{model})

패치 임베딩이 출력됨

분류기

BERT의 [class] 토큰처럼 학습가능한 임베딩을 임베딩 패치 시퀀스 추가 ($z_0^0 = x_{class}$)

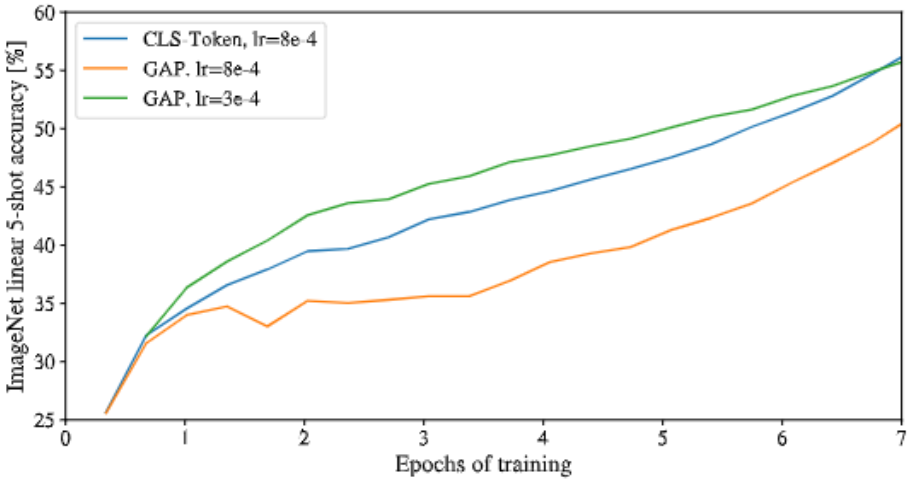
z_L^0 : 최종 L번째 레이어의 0번째 토큰. 트랜스포머 인코더의 output. 이미지 representation y로 사용됨

▼ BERT와의 유사성

- BERT에서도 "CLS" 토큰이 없다면 문장 내의 단어끼리의 관계는 잘 학습되지만, 문장 전체를 **대표**하는 요약된 정보를 얻기 어렵기에 텍스트 분류 작업에서 성능이 떨어짐
  - CLS 토큰은 항상 가장 앞에 위치 → 위치에 의해 영향을 받을 수 있는 추가적인 정보를 포함하지 않음. 문장 전체의 맥락 정보만을 저장

▼ CNN과의 차이점

- CNN 은 global average pooling이나 fully connected layer가 전체 이미지를 요약하므로 class token을 필요로하지 않
- 논문에서도 class token을 사용하지 않고 이미지 임베딩만 가지고 GAP를 사용하여 분류를 수행



사전학습과 파인튜닝 모두에서 분류기는 z_L^0를 사용

사전학습에서는 한 개의 히든 레이어를 가진 MLP, 파인튜닝에서는 하나의 선형 레이어

포지션 임베딩

위치 정보를 재학습하기 위해 패치 임베딩에 추가 됨.

각 패치가 이미지에서 어떤 위치에 있었는지 알려줌

일반적인 1D 사용. 2D로 해도 큰 성능 차이 없음

최종 임베딩 벡터 시퀀스는 인코더의 입력으로 사용됨

트랜스포머 인코더는 MHA, MLP 블록으로 이루어져있고 LN이 블록 이전, 잔차가 블록 이후에 더해진 형태

The MLP contains two layers with a GELU non-linearity.

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

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

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

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

Inductive bias

> 트랜스포머 구조에서 translation equivariance를 구현하는 것이 이 논문의 목적

cnn보다 훨씬 적은 inductive bias 가지고 있음

cnn: locality, 2d 이웃 구조, translation equivariance이 전체 모델의 각 레이어에 내재되어 있음

vit: mlp 레이어만 locality, translation equivariance 가짐. self-attention 레이어는 전역적

: 포지션 임베딩은 패치의 2d 위치 정보 가지지 않음/패치들 사이의 모든 위치 관계는 스크래치부터 학습됨

: 2d 이웃 구조는 매우 희소함. 패치로 나눌 때, 파인튜닝에서 포지션 임베딩 적용할 때

Hybrid Architecture

> cnn + mlp 기반의 classifier 모델 구조

raw 이미지 패치에 대한 대안으로 cnn의 feature map을 입력 시퀀스로 사용할 수 있음

패치 임베딩 projection E는 cnn feature map에서 추출된 패치에 적용됨

특별한 케이스로 해당 패치는 1x1의 크기를 가질 수 있음 = feature map의 공간적 차원을 flatten하고 트랜스포머 차원(D)으로 projection하여 만들어짐

FINE-TUNING AND HIGHER RESOLUTION

사전학습: 큰 데이터셋, 파인튜닝: 작은 downstream task

사전학습된 predictoin head 없애고 zero-initialized DxK feedforward 레이어 추가

사전 학습 때보다 더 높은 해상도로 파인튜닝하는 게 더 좋음

패치 사이즈는 동일하게 유지

메모리 사이즈에 따라 어떤 시퀀스 길이든 처리할 수 있음

EXPERIMENTS

ResNet, Vision Transformer (ViT), hybrid 비교

각 모델에 대한 데이터 필요량을 알아내기 위해 다양한 사이즈의 데이터로 사전학습후 다양한 벤치마크로 평가

SETUP

Datasets

학습데이터

- ImageNet-1k(1.3 M)
- ImageNet-21k(14 M)
- JFT-18k(303M)

벤치마크

- 19-Task VTAB 평가(Evaluate) - 적은 데이터셋을 활용한 Transfer Learning 성능 평가
- Natural : Pet, CIFAR
- Specialized : Medical이나 위성 이미지
- Structured : 기하학(Geometric)과 지역적(Localization)인 이해가 필요

Model Variants

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Table 1: Details of Vision Transformer model variants.

ViT: BERT를 베이스로 사용(Base, Large, Huge)

CNN(**BiT**): ResNet을 베이스로 사용. Group Normalization & Weight Standardization 사용→ transfer 향상

Hybrid

Training & Fine-tuning

학습 하이퍼파라미터: Optimizer - Adam, Batch Size - 4096, Weight Decay - 0.1

파인튜닝 하이퍼파라미터: Optimizer - SGD, Batch Size - 512, 해상도 변경 : 1. ViT-L/16 - 512 2. ViT-H/14 - 518

Metrics

Few-shot 정확도 : Regularized Least-squares Regression으로 측정

실시간 평가를 위해, Fine-Tunning 비용이 큰 경우 사용

Training Image의 (Frozen) Representation 추출 → {-1, 1} K Target Vector로 매핑하여 측정

COMPARISON TO STATE OF THE ART

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

Table 2: Comparison with state of the art on popular image classification benchmarks. We report mean and standard deviation of the accuracies, averaged over three fine-tuning runs. Vision Transformer models pre-trained on the JFT-300M dataset outperform ResNet-based baselines on all datasets, while taking substantially less computational resources to pre-train. ViT pre-trained on the smaller public ImageNet-21k dataset performs well too. *Slightly improved 88.5% result reported in Touvron et al. (2020).

noisy student: ImageNet에서 현재 sota인 모델

Semi-Supervised Learning을 적용한 Large EfficientNet

ViT 뒤의 숫자는 패치 크기

vit가 전반적으로 성능이 더 좋음

vit-l/16은 resnet과 거의 비슷한 성능

vit-h/14는 성능 크게 좋음

vit 모델이 계산량/학습시간 확연히 적음

ImageNet에 대한 few-shot 결과와 VTAB라는 적은 데이터에 대한 결과는 vit가 적은 데이터 전이에 유용하다는 것을 보여줌

PRE-TRAINING DATA REQUIREMENTS

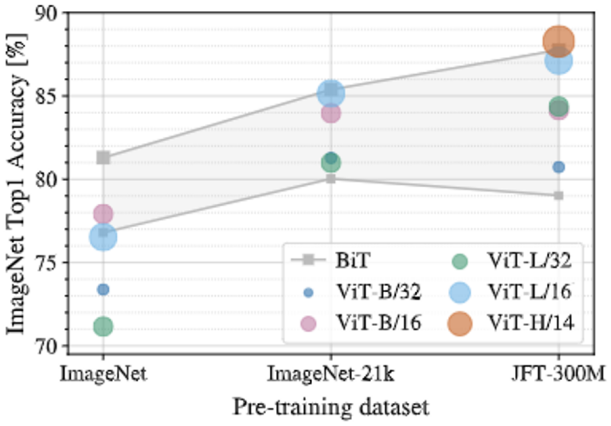


Figure 3: Transfer to ImageNet. While large ViT models perform worse than BiT ResNets (shaded area) when pre-trained on small datasets, they shine when pre-trained on larger datasets. Similarly, larger ViT variants overtake smaller ones as the dataset grows.

작은 데이터셋(=ImageNet)으로 사전학습 시 large vit는 bit보다 안 좋은 성능. 큰 데이터로 사전학습 해야 좋음

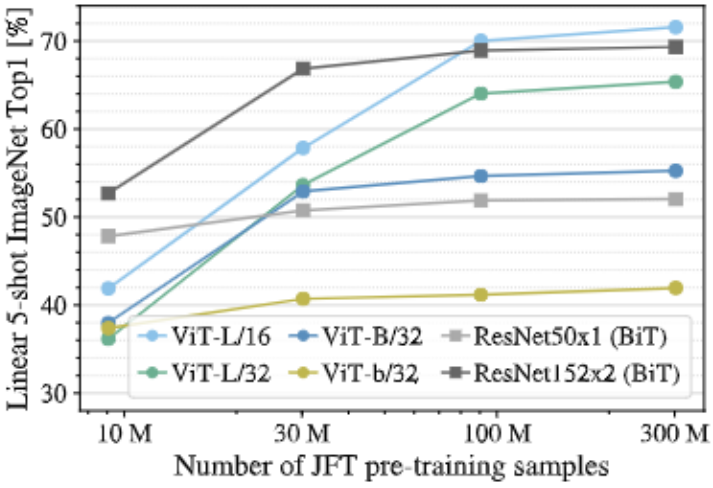


Figure 4: Linear few-shot evaluation on ImageNet versus pre-training size. ResNets perform better with smaller pre-training datasets but plateau sooner than ViT, which performs better with larger pre-training. ViT-b is ViT-B with all hidden dimensions halved.

다양한 사전학습 데이터셋 크기에 따른 few-shot evaluation 결과

resnet은 더 작은 사전학습 데이터셋에서 더 좋은 성능을 내지만 성능 향상이 더딤

vit-b와 vit-B는 히든레이어 차원 절반인 모델

> vit는 큰 데이터셋으로 사전학습 하는 게 좋다

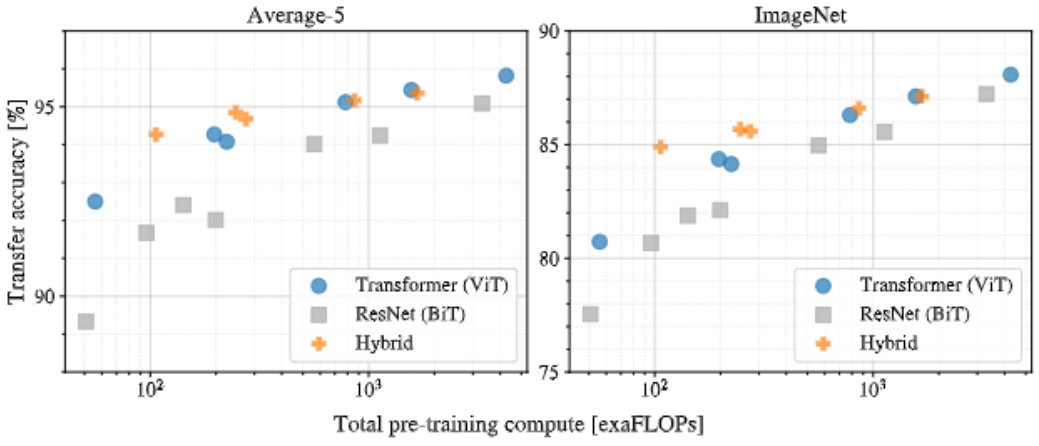


Figure 5: Performance versus pre-training compute for different architectures: Vision Transformers, ResNets, and hybrids. Vision Transformers generally outperform ResNets with the same computational budget. Hybrids improve upon pure Transformers for smaller model sizes, but the gap vanishes for larger models.

각 구조 별 연산량에 따른 전이학습 정확도 비교

resnet 보다는 항상 좋고, 작은 크기에서는 hybrid 보다 못하지만 큰 사이즈에서는 hybrid 뛰어 넘음

vit 성능 더 좋아질 수 있음(아직 포화 안 일어남)

INSPECTING VISION TRANSFORMER

internal representation 확인

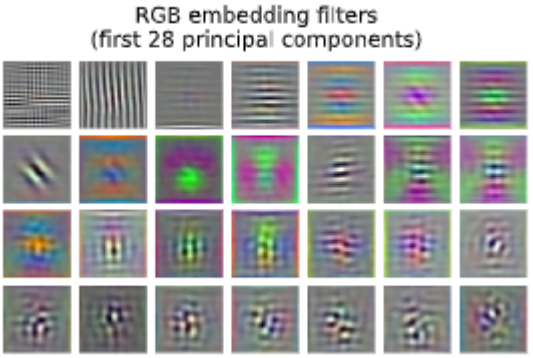
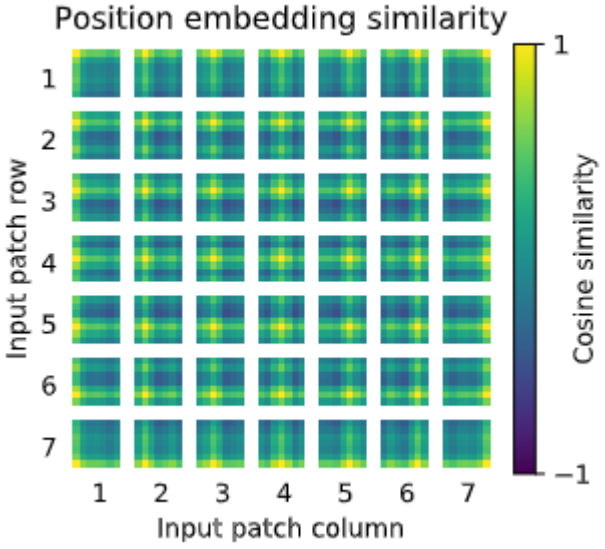


Figure 7: Left: Filters of the initial linear embedding of RGB values of ViT-L/32

ViT의 첫번째 레이어는 flatten된 패치를 낮은 차원으로 선형 projection → 학습된 임베딩 필터의 주요 구성요소를 보여줌



Center: Similarity 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.

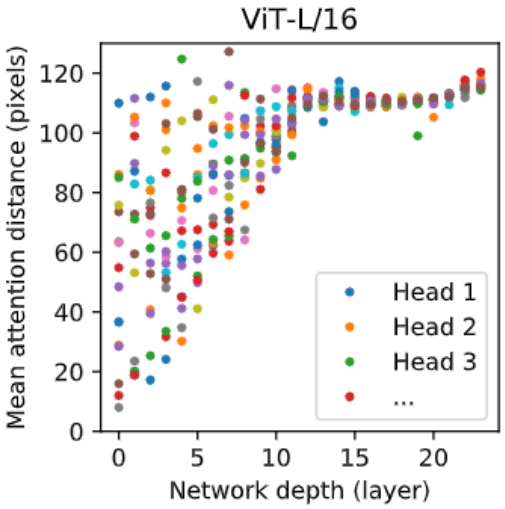
투영 이후에 학습된 포지션 임베딩은 패치 임베딩에 더해짐

모델이 위치 임베딩의 유사성을 통해 이미지 내의 거리를 인코딩하는 방법을 배운다는 것을 보여줌 = **더 가까운 패치들은 더 유사한 위치 임베딩을 가짐**

행-열 구조가 나타남 = 같은 행/열에 위치한 패치는 비슷한 임베딩을 가짐

sinusoidal 구조는 큰 그리드에서 더 잘 나타남

포지션 임베딩이 2d 이미지의 공간적 관계를 학습함 = 2d 임베딩이 더 나은 결과를 보이지 못하는 이유



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

self attention은 vit가 전체 이미지의 정보를 통합할 수 있도록 함(가장 낮은 레이어에서도)

어텐션 가중치에 따라 정보가 통합되어 있는 이미지 공간 내의 평균 거리를 계산하여 그 정도 확인

이 'attention distance'는 receptive field와 유사한 개념

몇몇 헤드는 이미지 대부분에 주목하는 것을 발견

다른 어텐션 헤드는 낮은 층에서 일관되게 작은 attention distance를 가짐 = highly localized attention

hybrid 모델에서는 덜 나타남

초기 합성곱 층과 유사한 기능을 수행할 수도 있음

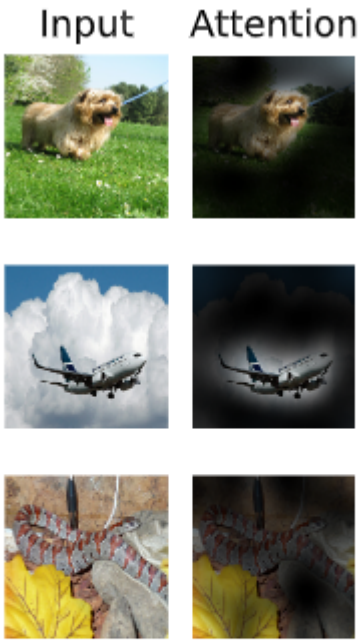


Figure 6: Representative examples of attention from the output token to the input space. See Appendix D.7 for details.

전반적으로 모델이 분류와 의미적으로 연관된 이미지 영역에 주의를 기울임

SELF-SUPERVISION

트랜스포머의 성공은 scalability와 large scale self-supervised pre-training에 기반함

self-supervision을 위해 masked patch prediction 수행(BERT와 비슷)

ViT-B/16 모델은 ImageNet에서 79.9%의 정확도

처음부터 훈련하는 것에 비해 2% 개선

supervised pre-training에 비해 4% 뒤쳐짐

이는 future work로

> Q. 왜 큰 데이터셋이서만 ViT 모델이 좋은 성능을 보일까?
>
> 가설1. Transformer가 CNN이 가진 translation equivariance 및 locality와 같은 **Inductive Bias를** 가지고 있지 않기 때문. 따라서 충

분한 데이터 없이는 일반화 성능이 떨어짐

▼ **inductive bias**

CNN의 유도 편향은 강하다 → 초기 지식과 가정이 강하다 → 내재된 복잡한 패턴을 학습이 어렵다

ViT의 유도 편향은 약하다 → 필요한 편향을 학습한다→ 패치 간의 관계나 이미지의 전반적인 구조를 학습한다 → 복잡한 패턴이나 글로벌한 정보를 잘 포착한다

CNN Inductive Bias

- **Locality (지역성) :**

  CNN은 이미지가 여러 작은 부분으로 구성되어 있고, 이 부분들에서 중요한 패턴(엣지, 텍스처)가 나타난다고 가정한다.

- **Translation Equivariance (전이 불변성) :**

  CNN은 이미지에서 어떤 패턴이 나타나는 위치가 조금 바뀌어도 여전히 그 패턴을 인식할 수 있다.

- **2D Neighborhood Structure (2D 이웃 구조) :**

  CNN은 이미지의 픽셀들이 2D 격자로 배치되어 있으며, 이웃한 픽셀들이 서로 밀접한 관계를 가진다고 가정한다.

ViT Inductive Bias

MSA: Global / MLP: Local

- **Locality (지역성) :**

  MLP에서만 지역성을 활용, ViT는 모든 patch들이 서로 동일한 중요도를 가지며 patch들 간의 관계를 학습을 통해 알아낸다.

- **Translation Equivariance (전이 불변성) :**

MLP에서만 지역성을 활용, patch의 위치 정보는 별도로 학습해야 한다.

- **2D Neighborhood Structure (2D 이웃 구조) :**

  처음 image를 patch로 나눌 때, 위치 임베딩을 통해 패치의 위치 정보를 제공할 때 사용하지만, patch들 간의 2D 이웃 구조를 중요하게 여기지 않는다. 즉, 이웃한 픽셀간 관계에 강한 가정이 없다.

가설2. cnn에 비해 “전체” 이미지를 서로 서로 비교하기 때문에 이미지 분석을 위한 적절한 성능에 도달하기까지 더 많은 데이터셋이 필요함

figure 7 right에서도 볼 수 있음