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

2024.3.27

Hae-Yeon Kim

Division of AI & Computer Engineering

Abstract

- Transformer architecture applied directly to **sequences of image patches can perform very well on image classification tasks.**
- **When pre-trained on large amounts of data** and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.

Introduction

- we experiment with applying a standard Transformer directly to images, with the fewest possible modifications.
- we **split an image into patches and provide the sequence of linear embeddings of these patches as an input to a Transformer.** Image patches are treated the same way as tokens (words) in an NLP application.
- **Large scale training trumps inductive bias.**

Inductive bias

- Transformer는 CNN에 고유한 inductive biases이 부족하므로 충분하지 못한 양의 데이터으로 학습할 때 일반화가 잘 되지 않는다.
- 머신러닝 알고리즘은 타겟 함수를 학습하고 학습 데이터를 넘어 일반화하기 위해 모델에 의해 가정된 Inductive Bias(유도 편향)를 의도적으로 사용한다
- 머신러닝 문제를 더 잘 풀기 위해 사전 정보를 통해 추가된 가정을 Inductive bias라고 할 수 있다.
- 특정 데이터셋에 대해 더 좋은 성능을 얻고자 Inductive bias를 의도적으로 강제해준다.

■ Inductive bias

Component	Entities	Relations	Rel. inductive bias	Invariance
Fully connected	Units	All-to-all	Weak	-
Convolutional	Grid elements	Local	Locality	Spatial translation
Recurrent	Timesteps	Sequential	Sequentiality	Time translation
Graph network	Nodes	Edges	Arbitrary	Node, edge permutations

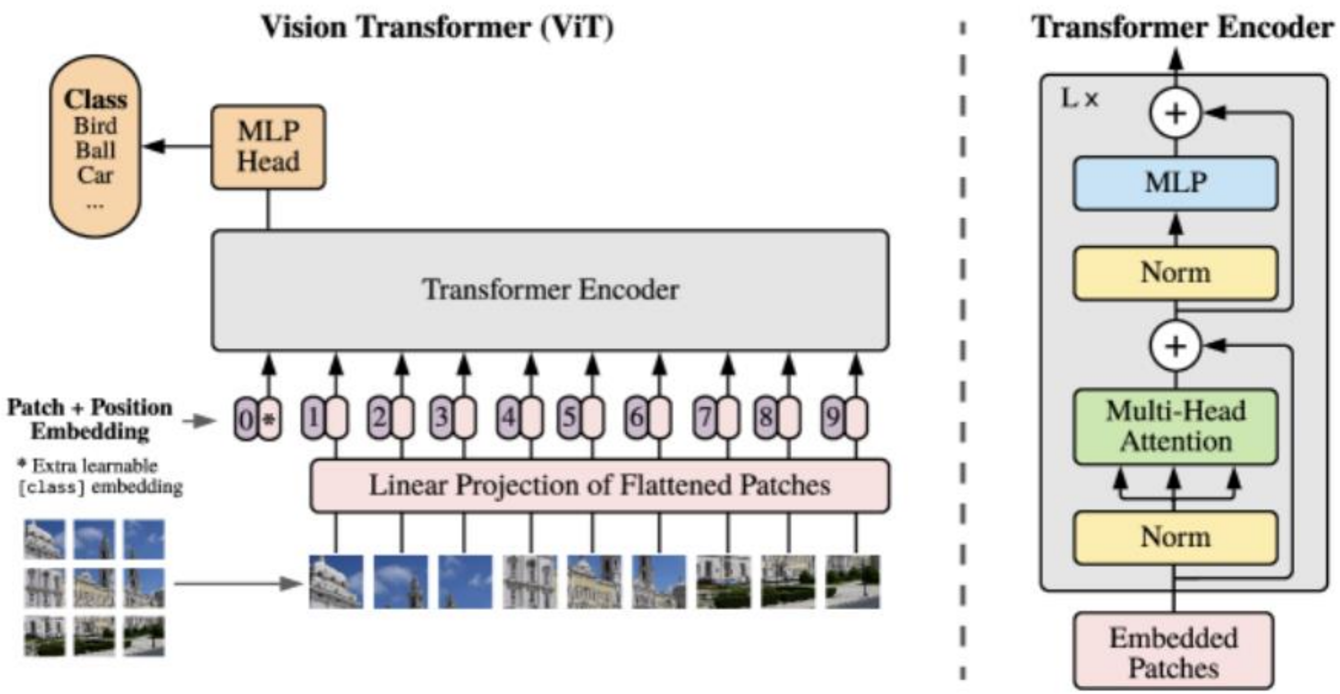
Table 1: Various relational inductive biases in standard deep learning components. See also Section 2.

- Inductive bias is the set of assumptions that a machine learning algorithm makes about the relationship between input variables (features) and output variables (labels) based on the training data.

- Better than other work
  - large scale pre-training makes vanilla transformers
  - Cordonnier et al. (2020) use a small patch size of 2×2 pixels, small-resolution images, while we handle medium-resolution images as well.

  - 데이터셋 사이즈에 따른 성능 연구 트랜스포머를 훈련

■ Structure


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 * 1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches
Transformer Encoder
L x
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches

■ Step

1. Split an image into patches (fixed P*P)
2. Flatten the patches and Tokenization (N = HW / P^2)

$$\mathbf{x} \in \mathbb{R}^{H \times W \times C} \qquad \mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$$

3. Produce lower-dimensional linear embeddings from the flattened patches
4. Add positional embeddings
5. Feed the sequence as an input to a standard transformer encoder
6. Pretrain the model with image labels (fully supervised on a huge dataset)
7. Finetune on the downstream dataset for image classification

- ViT is pretrained on the large dataset and then fine-tuned to small ones.
  - The only modification is to discard the prediction head (MLP head) and attach a new $D \times K$ linear layer, where K is the number of classes of the small dataset.

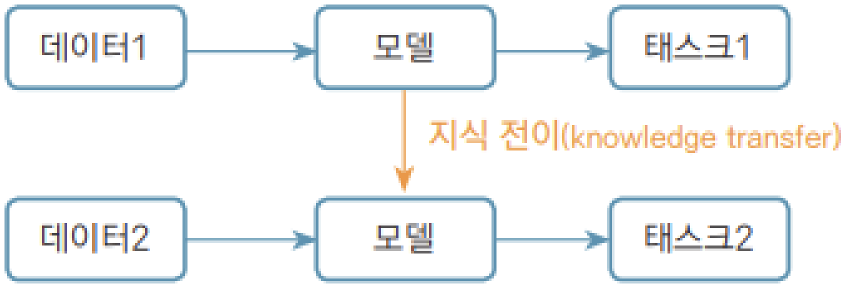


그림 1-6 트랜스퍼 러닝 개념도

▪ Comparison to State of the Art

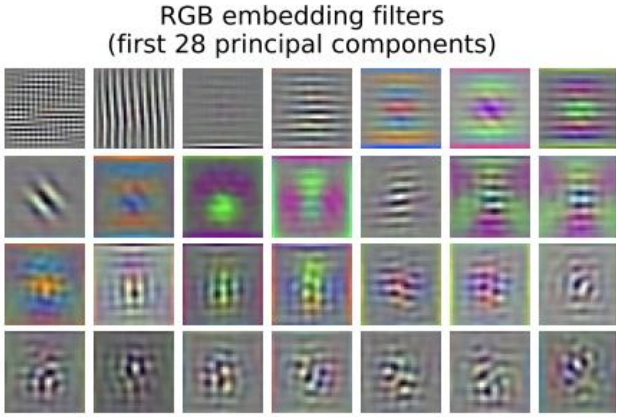
Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Table 1:Details of Vision Transformer model variants.

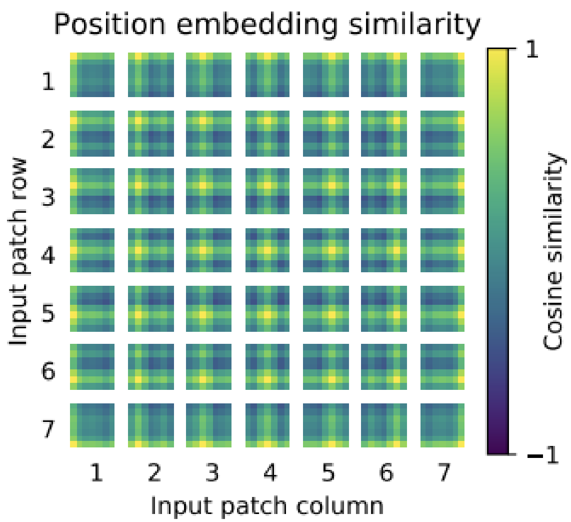
	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4 / 88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 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.

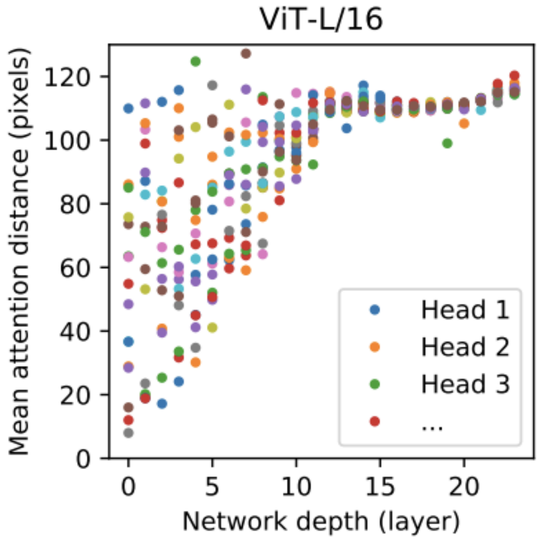
▪ Inspecting Vision Transformer



< fig1. Embedding projection>



< fig2. Position embedding>



< fig3. Self attention>