Lab Meeting

2024.09.16

Department of Defense AI Convergence Engineering
Seoul National University of Science and Technology

Kim Min

AIR Lab
서울과학기술대학교 인공지능 / 환경원격탐사 연구실

Published as a conference paper at ICLR 2021

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

Alexey Dosovitskiy[*,†], **Lucas Beyer**[*], **Alexander Kolesnikov**[*], **Dirk Weissenborn**[*], **Xiaohua Zhai**[*], **Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby**[*,†]

[*]equal technical contribution, [†]equal advising

Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com

ABSTRACT

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer 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.[1]

- International Conference on Learning Representations 2021년에 발표

AIR Lab
서울과학기술대학교 인공지능 / 환경원격탐사 연구실

CONTENTS

1. Introduction

- Inductive bias
- Vision Transformer(ViT) 개요

1. Inductive Bias

- Inductive bias는 training에서 보지 못한 데이터 대해서도 적절한 귀납적 추론이 가능하도록 하기 위해 모델이 가지고 있는 가정들의 집합을 의미
  - 보지 못한 상황을 해결하기 위해 , **추가적인 가정**을 활용해서 문제를 해결
- DNN의 inductive bias 예시
  - ➢ Fully connected: 입력 및 출력 element가 모두 연결되어 있으므로 구조적으로 특별한 relational inductive bias를 가정하지 않음
  - ➢ Convolutional: CNN은 작은 크기의 kernel로 이미지를 지역적으로 보며, 동일한 kernel로 이미지 전체를 본다는 점에서 locality와 transitional invariance 특성을 가짐
  - ➢ Recurrent: RNN은 입력한 데이터들이 시간적 특성을 가지고 있다고 가정하므로 sequentiality와 temporal invariance 특성을 가짐
- Transformer는 CNN, RNN보다 상대적으로 낮은 inductive bias를 가짐

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

2. Vision Transformer (ViT) 개요

- 본 연구에서는 NLP에서 사용되는 standard Transformer를 이미지에 그대로 적용하여 이미지 분류에 좋은 성능을 도출한 Vision Transformer(ViT)를 제안

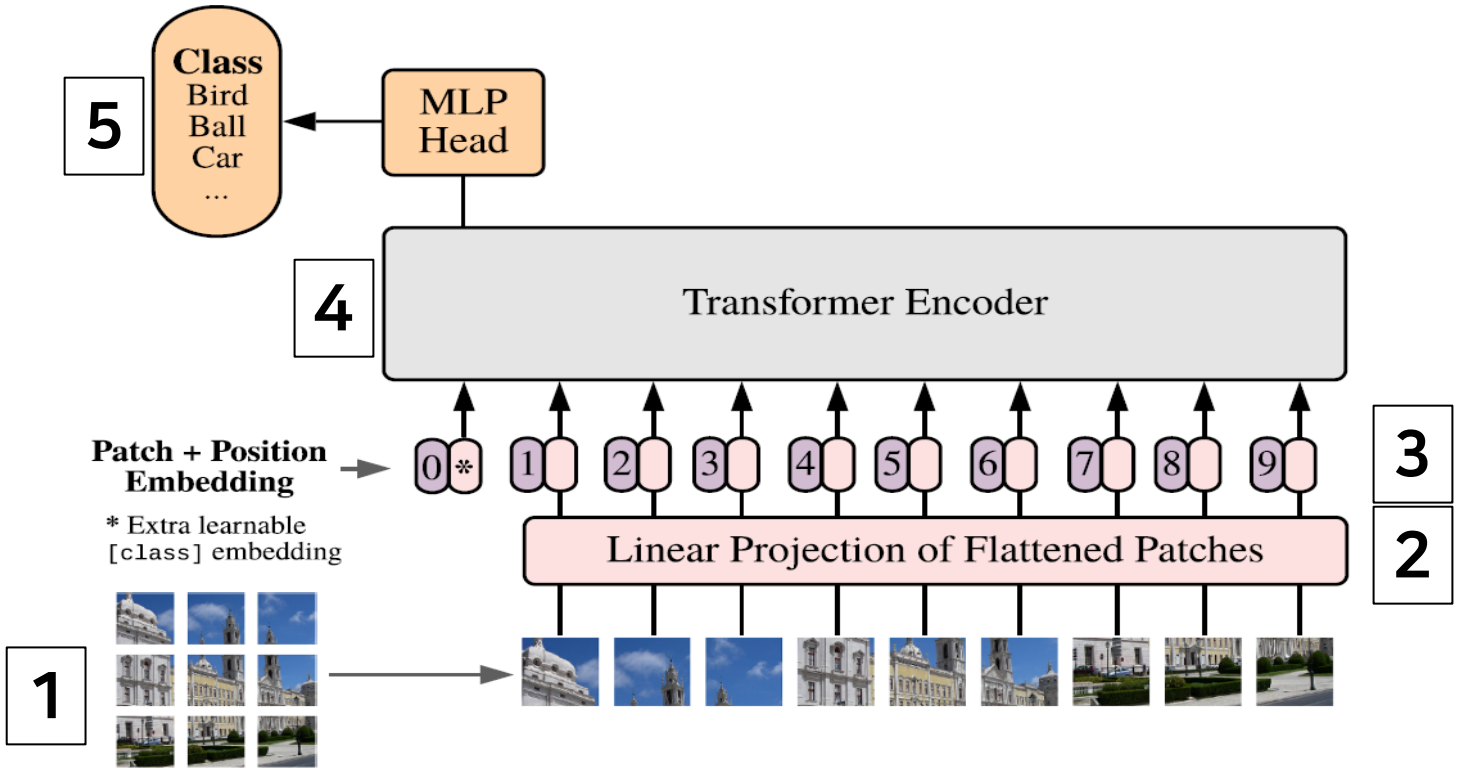
- 이미지를 여러 패치로 분할한 후, 이를 NLP의 단어로 취급하여 각 패치의 linear embedding을 순서대로 Transformer의 input으로 넣어 이미지를 분류

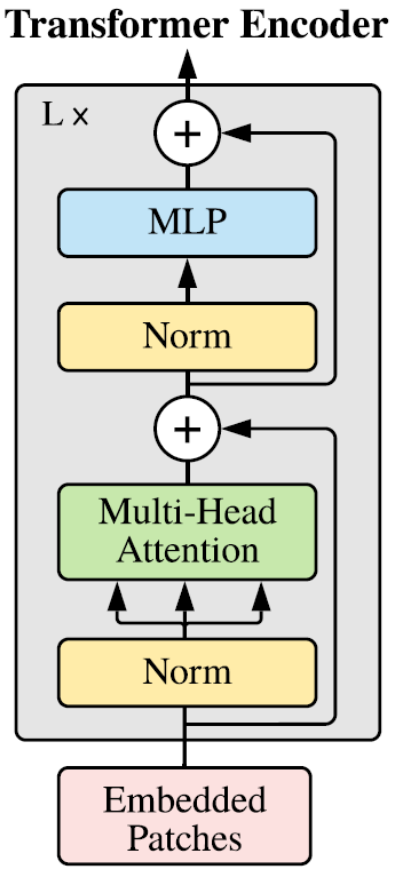
- Dataset size에 따른 성능 차이
  - Mid-sized dataset(ImageNet)에서는 ResNet보다 낮은 정확도로, CNN 보다 inductive bias가 낮아
  - Large-scale dataset에 pre-trained된 ViT를 transfer learning 했을 때, ViT가 SOTA 성능을 달성
    - 이를 통해, **large scale 학습이 낮은 inductive bias로 인한 성능 저하를 해소**시키는 것을 확인

2. Proposed Transformer

1. ViT 모델 구조

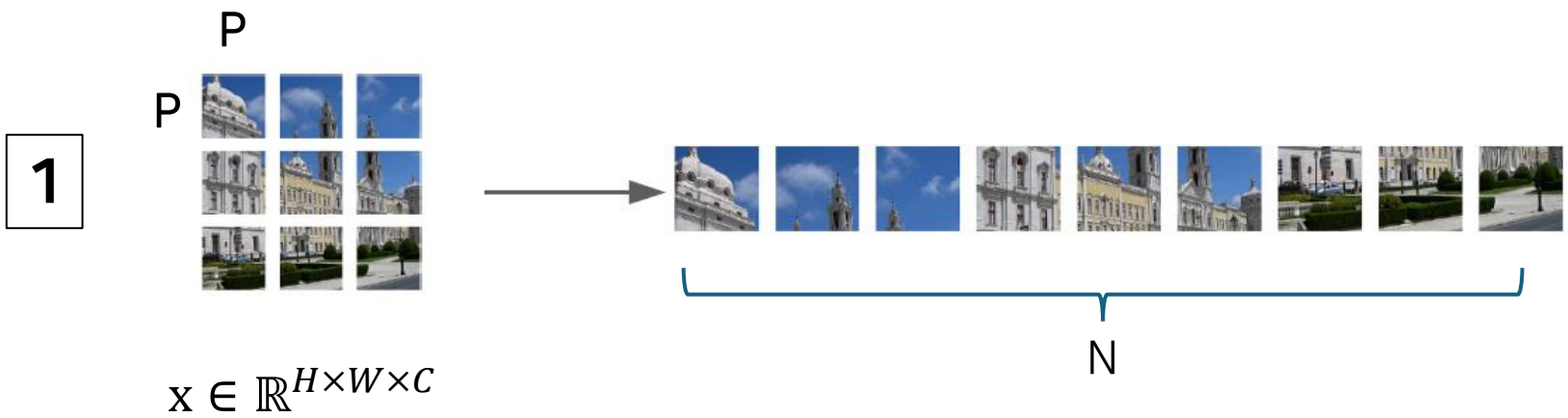
- ViT의 전체 구조





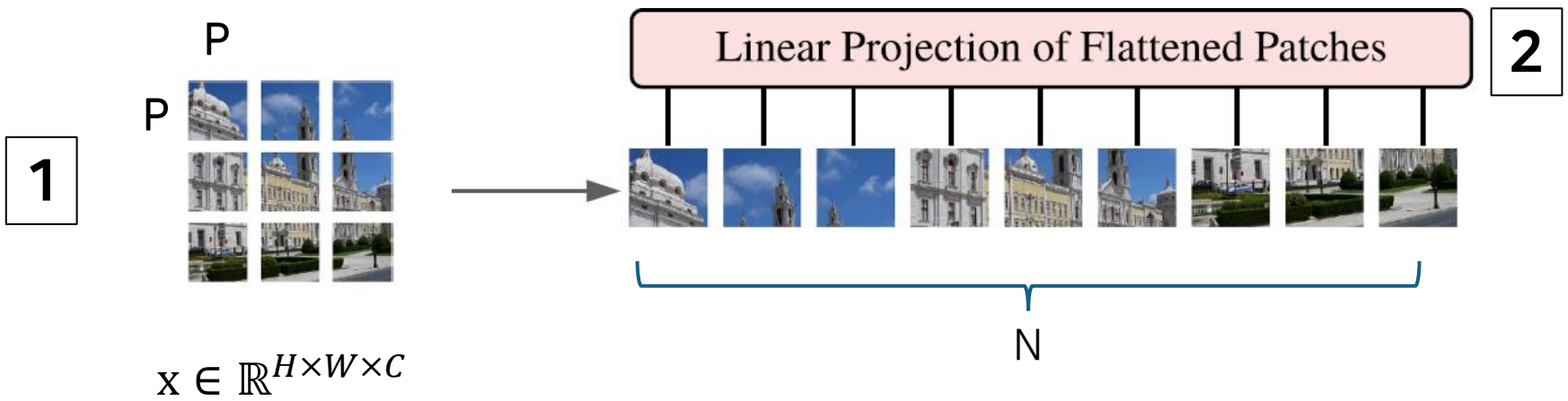
1. ViT 모델 구조

- ViT 동작 과정
  - Step 1. 이미지 $\mathrm{x} \in \mathbb{R}^{H \times W \times C}$가 있을 때, 이미지를 $(P \times P)$ 크기의 패치 N개로 분할하여 패치 sequence $\mathrm{x}_p \in \mathbb{R}^{N \times (P^2 \times C)}$를 구축함



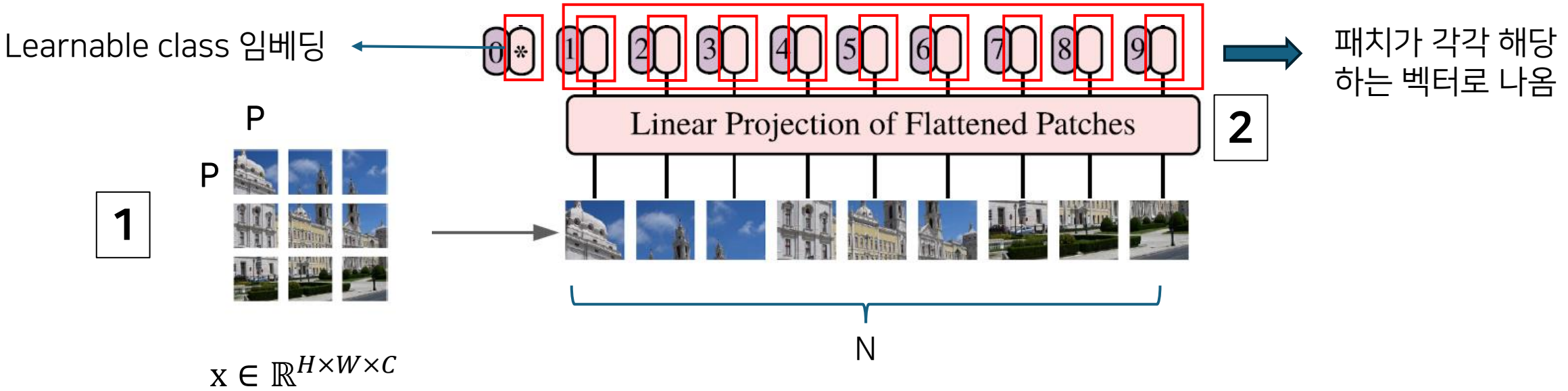
1. ViT 모델 구조

- ViT 동작 과정
  - Step 1. 이미지 $\mathrm{x} \in \mathbb{R}^{H \times W \times C}$가 있을 때, 이미지를 $(P \times P)$ 크기의 패치 N개로 분할하여 패치 sequence $\mathrm{x}_p \in \mathbb{R}^{N \times (P^2 \times C)}$를 구축함
  - Step 2. Trainable linear projection을 통해 x_p의 각 패치를 flatten한 벡터를 D차원으로 변환한 후, 이를 패치 임베딩으로 사용



1. ViT 모델 구조

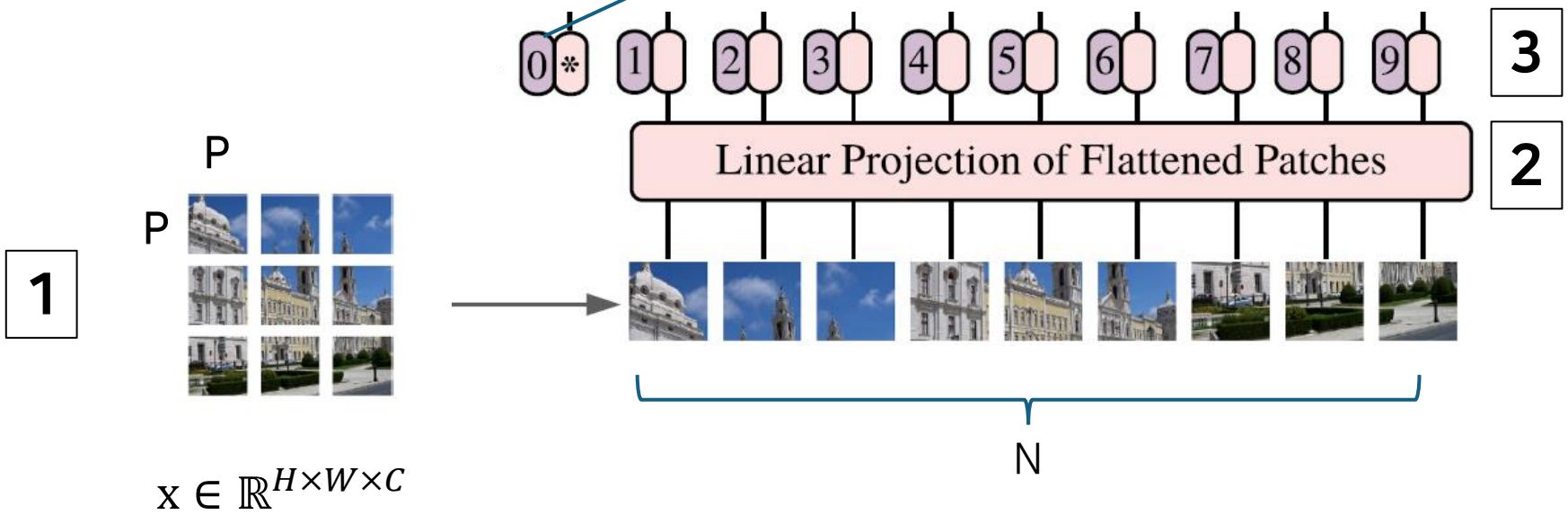
- ViT 동작 과정
  - Step 1. 이미지 $\mathrm{x} \in \mathbb{R}^{H \times W \times C}$가 있을 때, 이미지를 $(P \times P)$ 크기의 패치 N개로 분할하여 패치 sequence $\mathrm{x}_p \in \mathbb{R}^{N \times (P^2 \times C)}$를 구축함
  - Step 2. Trainable linear projection을 통해 x_p의 각 패치를 flatten한 벡터를 D차원으로 변환한 후, 이를 패치 임베딩으로 사용



1. ViT 모델 구조

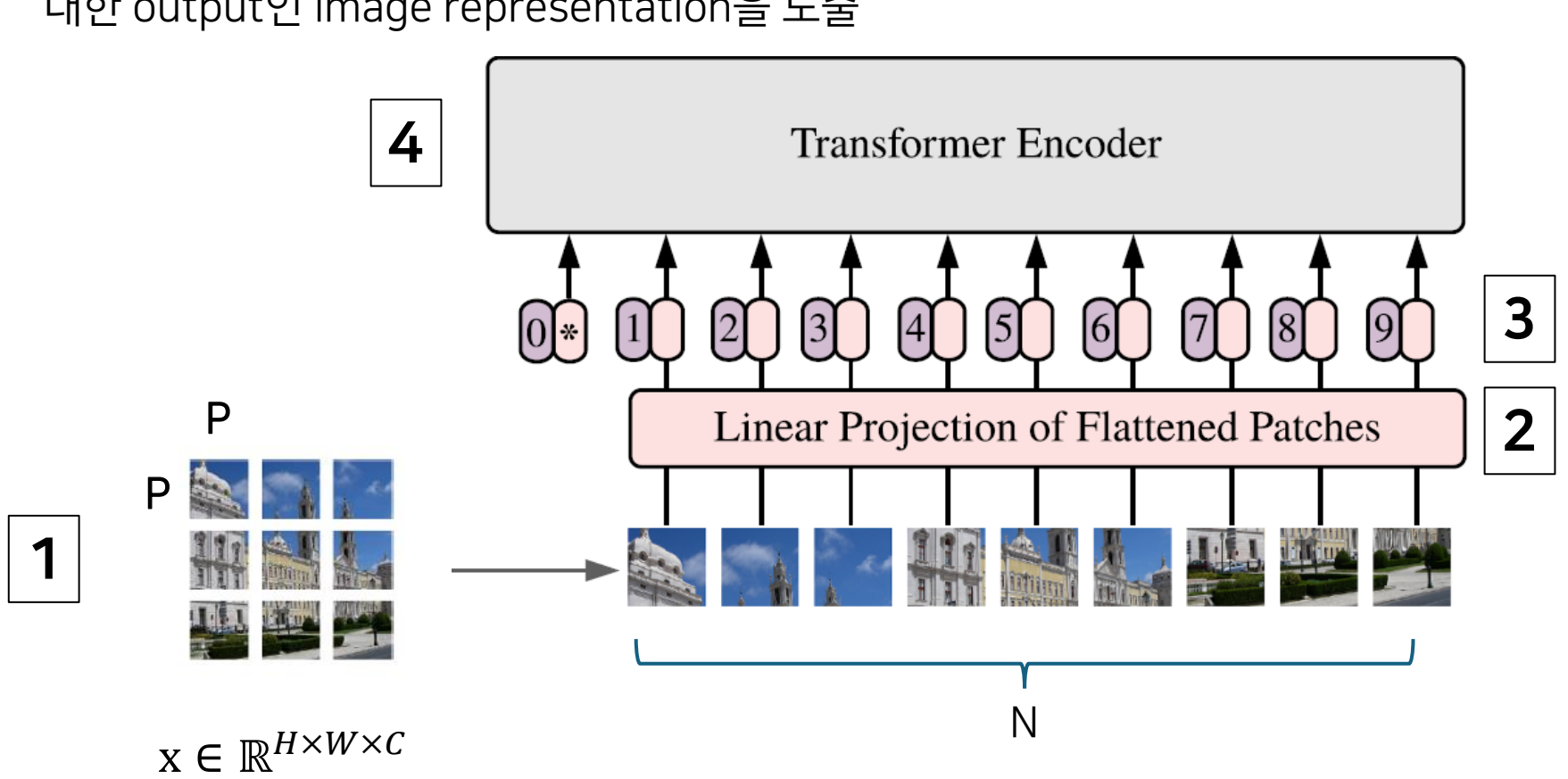
- ViT 동작 과정
  - Step 3. Learnable 임베딩과 패치 임베딩에 learnable position 임베딩을 더함

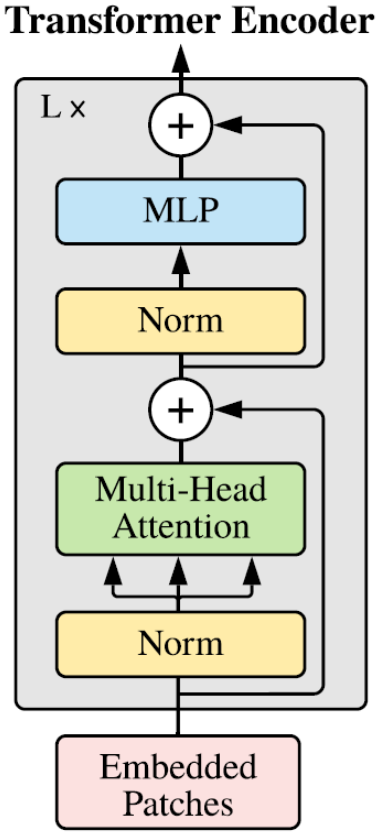
$$z_0 = [x_{class};\ x_p^1 E; x_p^2 E; \cdots x_p^N E] + E_{pos}, \quad E \in \mathbb{R}^{(P^2 \times C) \times D}, E \in \mathbb{R}^{(N+1) \times D}$$



1. ViT 모델 구조

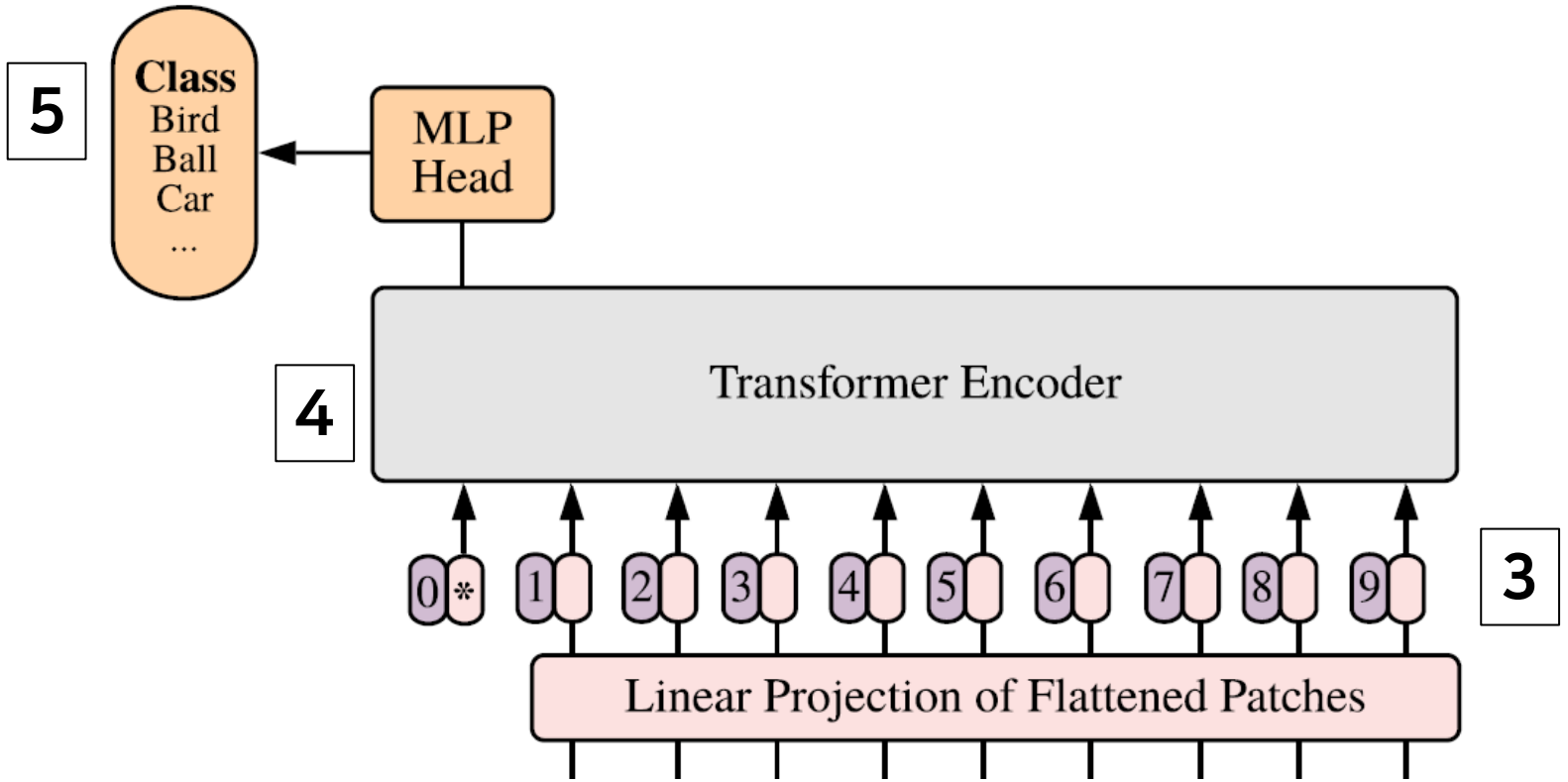
- ViT 동작 과정
  - Step 4. 임베딩을 Transformer encode에 input으로 넣어 마지막 layer에서 class embedding에 대한 output인 image representation을 도출





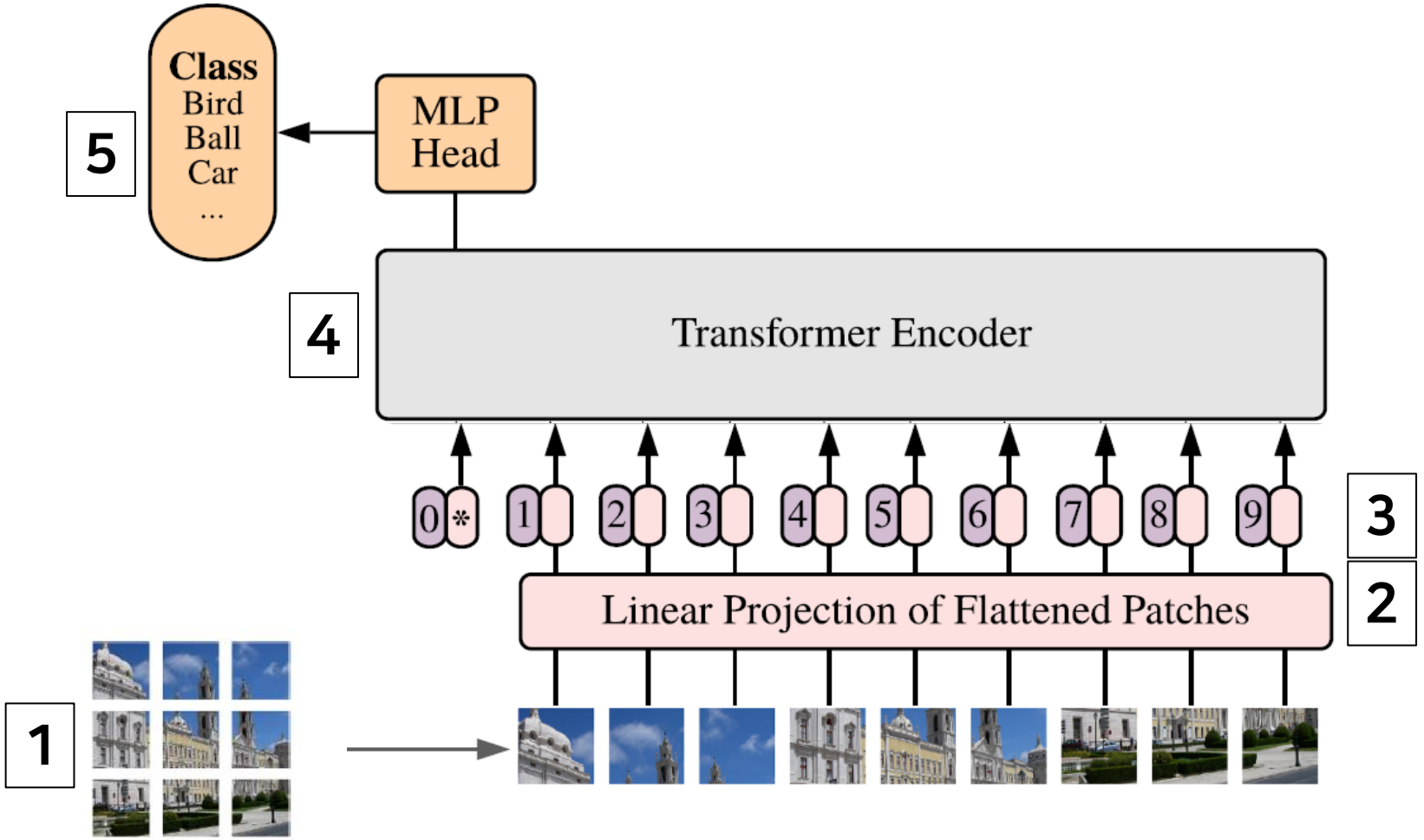
1. ViT 모델 구조

- ViT 동작 과정
  - Step 5. MLP에 image representation을 input으로 넣어 이미지의 class를 분류



1. ViT 모델 구조

- ViT 동작 과정 정리



2. Positional Embedding

- 본 논문에서는 1D position 임베딩을 ViT에 적용
  - ➢ No Pos. Embedding: 입력 패치들을 하나의 집합으로 간주
  - ➢ **1-D Pos. Embedding: 입력을 순차적으로 배치한 패치의 sequence로 간주**
  - ➢ 2-D Pos. Embedding: 입력을 2차원 그리드로 구성된 패치들로 간주하고 X축 Y축 각각에 대해 임베딩 학습
  - ➢ Relative Pos. Embedding: 절대적인 위치 대신, 패치 간의 상대적 거리를 사용하여 공간 정보를 인코딩

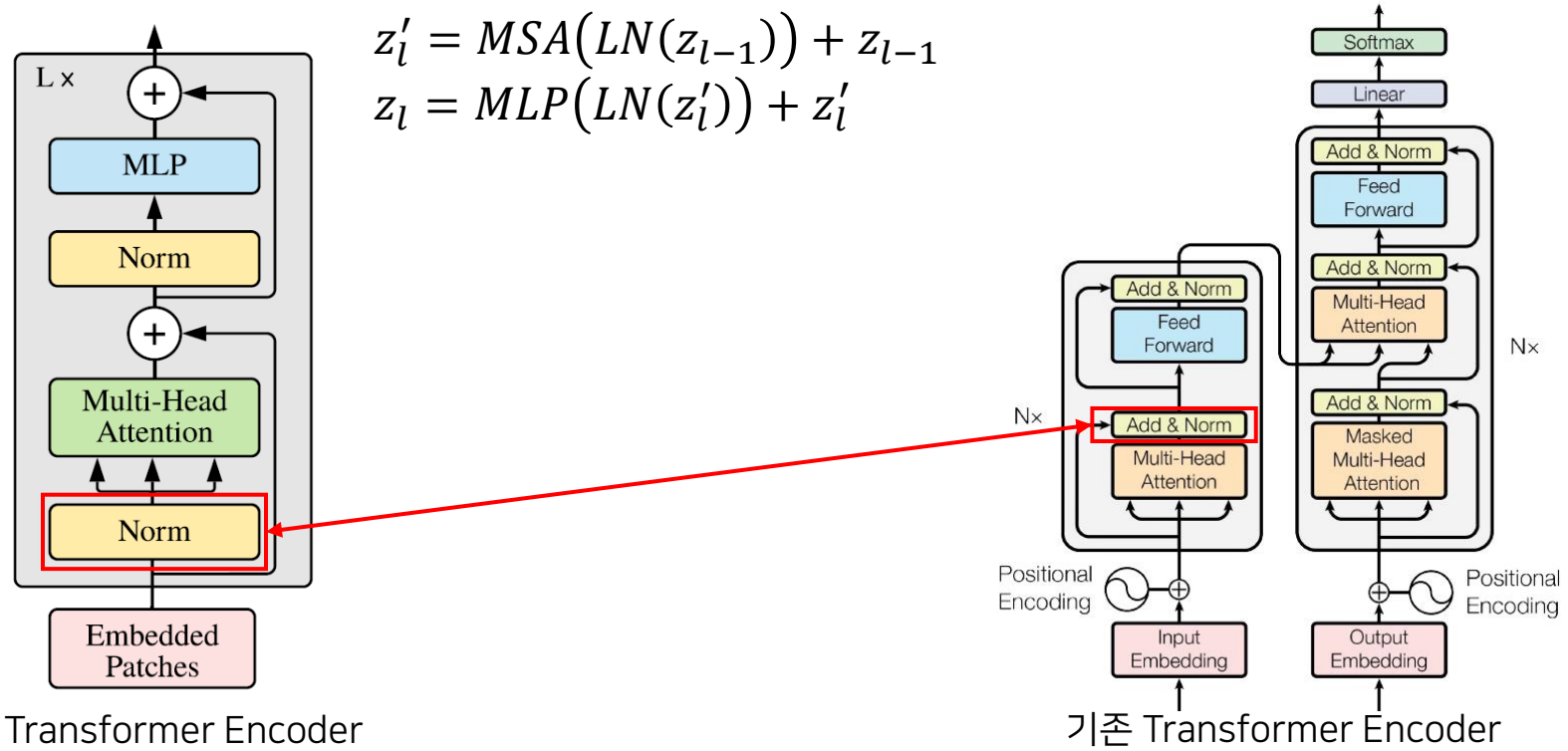
Pos. Emb.	Default/Stem	Every Layer	Every Layer-Shared
No Pos. Emb.	0.61382	N/A	N/A
1-D Pos. Emb.	0.64206	0.63964	0.64292
2-D Pos. Emb.	0.64001	0.64046	0.64022
Rel. Pos. Emb.	0.64032	N/A	N/A

Table 8: Results of the ablation study on positional embeddings with ViT-B/16 model evaluated on ImageNet 5-shot linear.

3. Transformer Encoder

- ViT는 Multi-head Self Attention(MSA)와 MLP block으로 구성
- MLP는 2개의 layer를 가지며, GELU activation function을 사용
- 각 block의 앞에는 Layer Norm(LN)을 적용하고, 각 block의 뒤에는 residual connection을 적용함

$$z'_l = MSA\left(LN(z_{l-1})\right) + z_{l-1}$$
$$z_l = MLP\left(LN(z'_l)\right) + z'_l$$

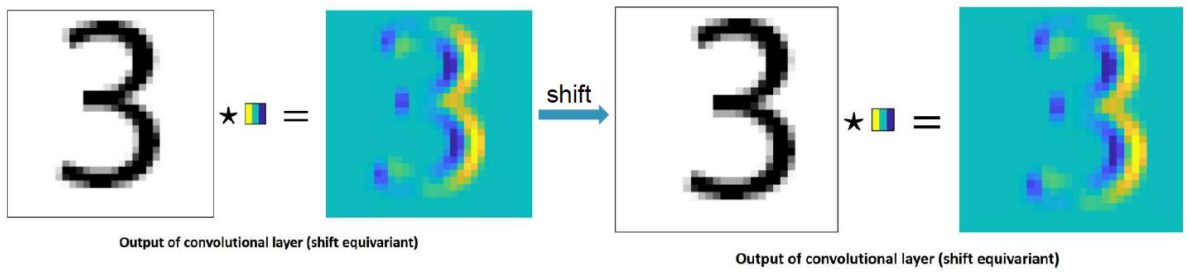


Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. 31st Conference on Neural Information Processing Systems.

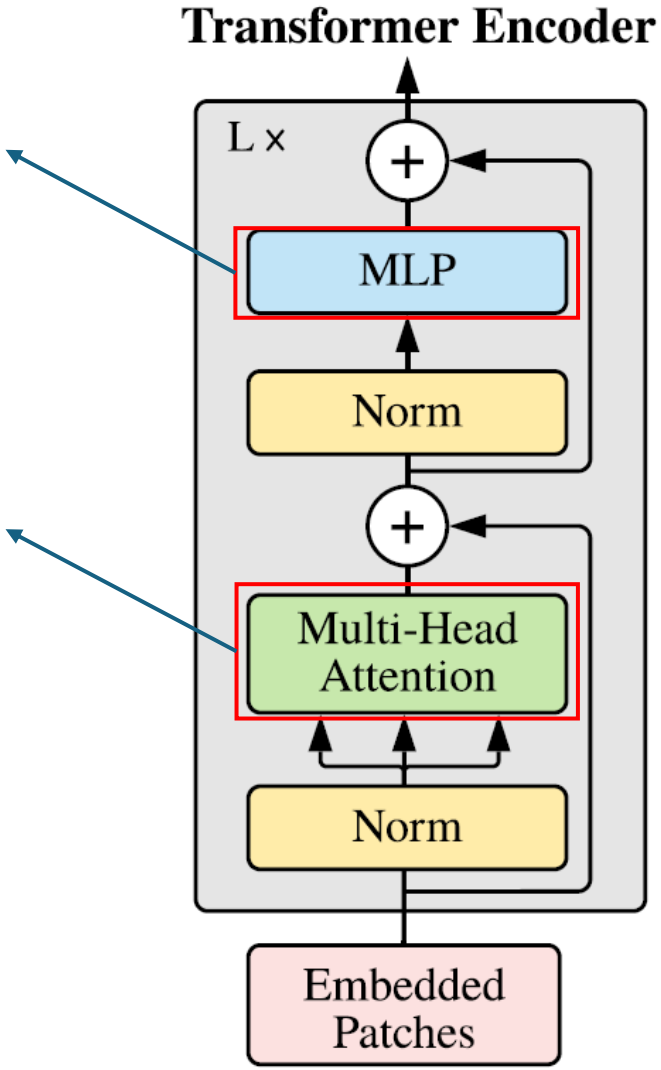
4. Inductive Bias

- ViT에서 MLP는 locality와 translation equivariance가 있음
  - ➢ MLP는 입력 데이터에 대해 각각 독립적이며, 각 뉴런은 특정한 입력 패치나 픽셀에 대해서만 계산(locality)

- MSA는 global하기 때문에 CNN보다 image-specific한 inductive bias가 낮음
  - ➢ Self-attention 메커니즘이 입력 데이터의 모든 요소가 서로 상호 작용할 수 있도록 설계

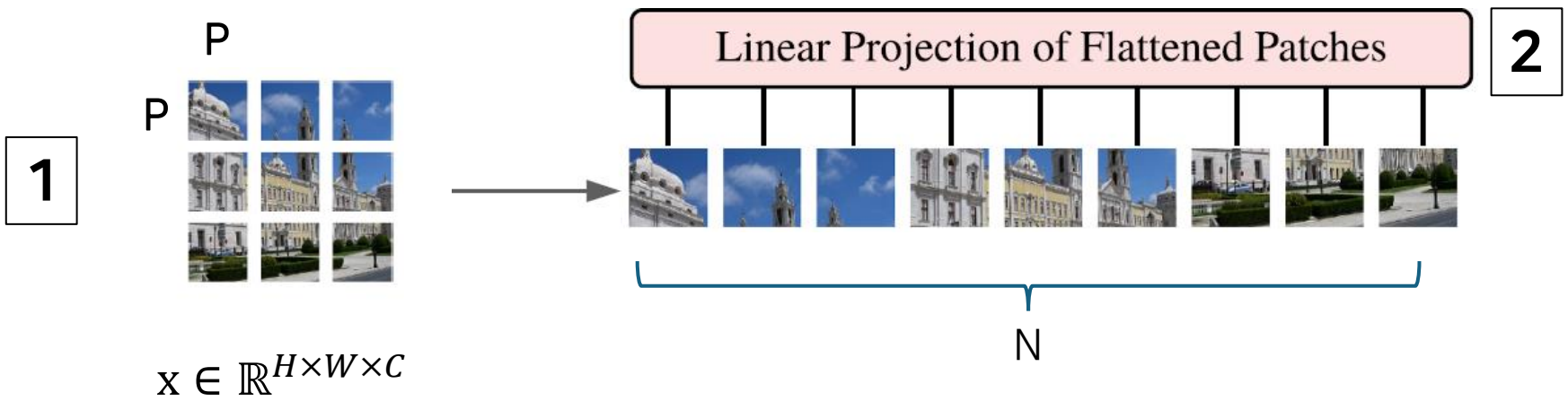


translation equivariance



4. Inductive Bias

- ViT에서는 모델에 아래 두가지 방법을 사용하여 inductive bias의 주입을 시도
  - ✓ 이미지를 패치로 나누는 과정

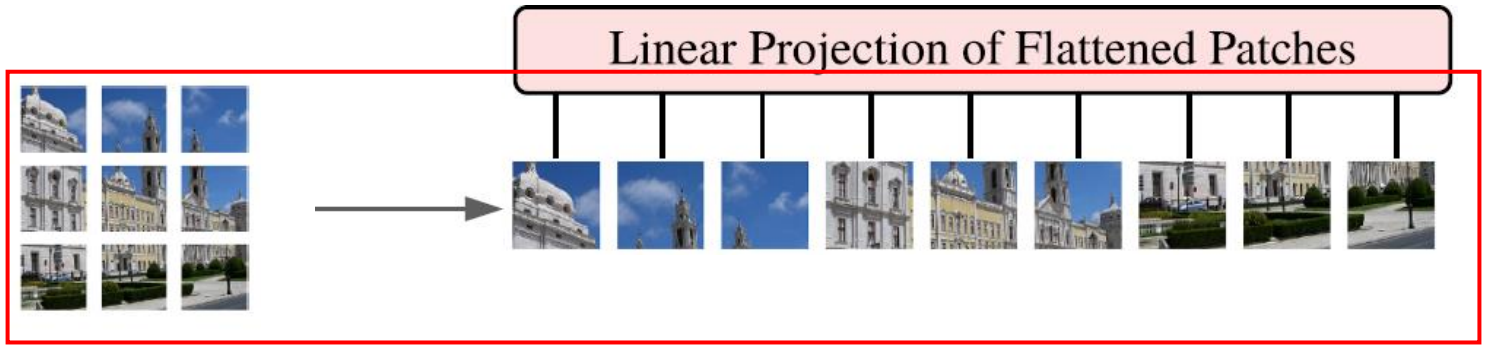


4. Inductive Bias

- ViT에서는 모델에 아래 두가지 방법을 사용하여 inductive bias의 주입을 시도
  - ✓ 이미지를 패치로 나누는 과정
  - ✓ fine-tuning에서 다른 해상도의 이미지에 대해 위치 임베딩을 조정

5. Hybrid Architecture

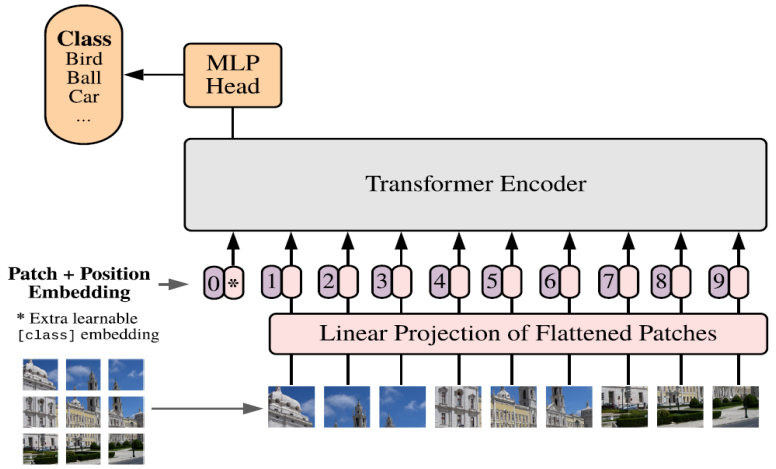
- ViT는 raw image가 아닌 CNN으로 추출한 raw image의 feature map을 활용하는 hybrid architecture로도 사용 가능
- Feature map은 이미 raw image의 공간적 정보를 포함하고 있으므로 hybrid architecture는 패치 크기를 1x1로 설정
- 1x1 크기의 패치를 사용할 경우 feature map의 공간 차원을 flatten하여 각 벡터에 linear projection을 적용



raw image 대신 feature map을 활용

6. Fine-tuning and Higher Resolution

- Large scale로 ViT를 pre-training한 후, 해당 모델을 downstream task에 fine-tuning하여 사용 가능
- ViT를 fine-tuning 할 때, ViT의 pre-trained prediction head를 zero-initialized feedforward layer로 대체함
- ViT를 fine-tuning 할 때, pre-training과 동일한 패치의 크기를 사용하기 때문에 고해상도 이미지로 fine-tuning을 하면 sequence 길이가 더 길어짐
- ViT는 가변적 길이의 패치들을 처리할 수 있지만, pre-trained position embedding은 의미가 사라지므로 pre-trained position embedding을 원본 이미지의 위치에 따라 2D interpolation하여 사용



3. Experiments

1. Dataset

- Class와 이미지의 개수가 각각 다른 3개의 데이터셋을 기반으로 pre-trained됨
- Downstream task에 따라 pre-trained ViT의 representation 성능을 검증
  - ✓ ReaL labels, CIFAR-10/100, Oxford-IIIT Pets, Oxford Flowers-102
  - ✓ 19-task VTAB classification suite

Pre-trained Dataset	The number of Classes	The number of images
ImageNet – 1k	1k	1.3M
ImageNet – 21k	21k	14M
JFT	18k	303M

2. Model Variants

- ViT는 아래와 같이 총 3개의 model variant에 대하여 실험을 진행하였으며, 다양한 패치 크기에 대해 실험을 진행
- Baseline CNN은 batch normalization layer를 group normalization으로 변경하고 standardized convolutional layer를 사용하여 transfer learning에 적합한 Big Transformer (BiT) 구조의 ResNet을 사용해서 ViT와 비교

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.

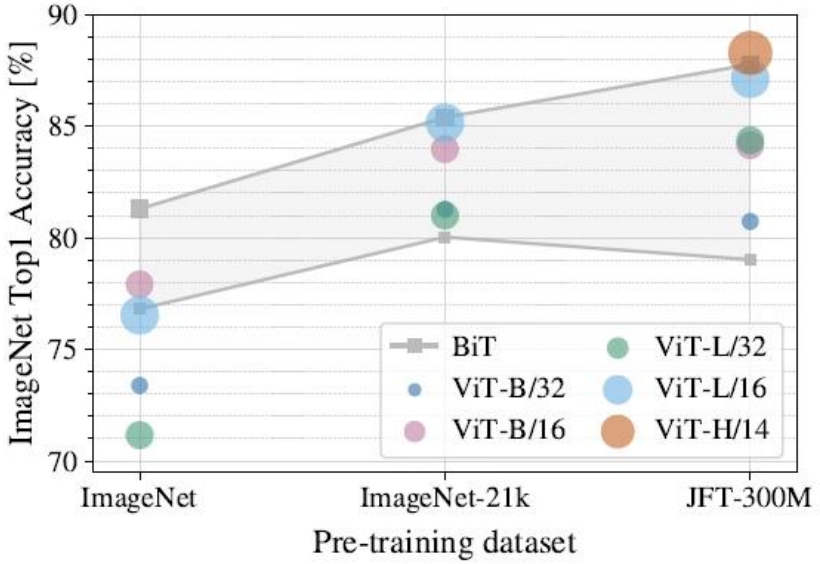
3. Results

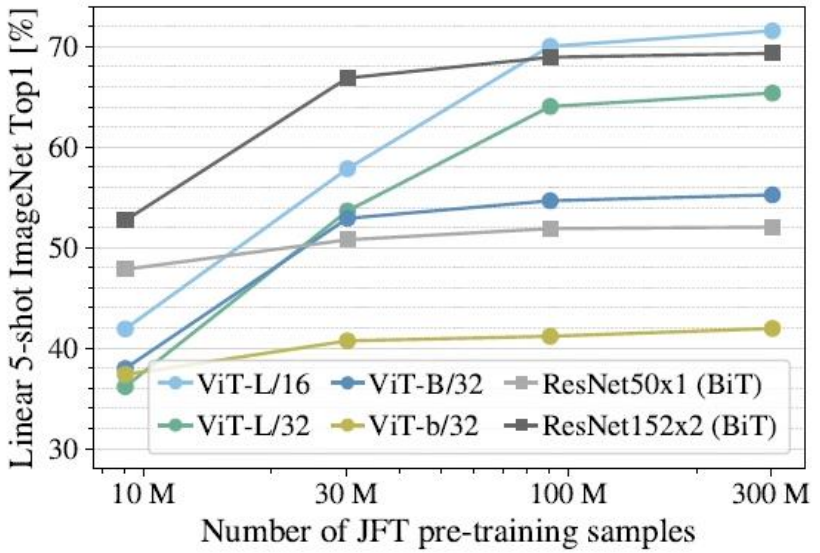
- 본 실험에서는 14 x 14 패치 크기를 사용한 Vit-Huge(ViT-H/14)와 16 x 16 패치 크기를 사용한 ViT-Large의 성능을 baseline과 비교
- JFT 데이터셋에서 pre-training한 ViT-L/16 모델이 모든 downstream task에 대하여 BiT-L보다 높은 성능을 보임
- ViT-L/14 모델은 ViT/16 모델보다 향상된 성능을 도출하였으며, BiT-L 모델보다 학습시간이 훨씬 짧음

	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

3. Results

- 본 실험에서는 pre-training 데이터셋의 크기에 따른 fine-tuning 성능을 확인함
- 각 데이터셋에 대하여 pre-training한 ViT를 ImageNet에 transfer learning한 정확도를 확인한 결과, 데이터가 클수록 ViT가 BiT보다 성능이 좋고 크기가 큰 ViT 모델이 효과가 있음
- JFT를 각각 다른 크기로 랜덤 샘플링한 데이터셋을 활용하여 실험을 진행한 결과, 작은 데이터셋에서 CNN의 inductive bias 효과가 있으나 큰 데이터셋에서는 데이터로부터 패턴을 학습하는 것만으로 충분





4. Conclusion

4. Conclusion

- Transformer를 image patch로 처리해 이미지 인식에 적용한 접근법이 Large-Scale dataset에서 CNN을 능가하는 성능을 보임

- Vision Transformer(ViT)는 사전 학습을 통해 image classification task에서 SOTA보다 비슷하거나 더 나은 결과를 보여줌

- 향후, Detection and Segmentation, self-supervised pre-training 에 적용 방법에 대한 추가 연구 필요

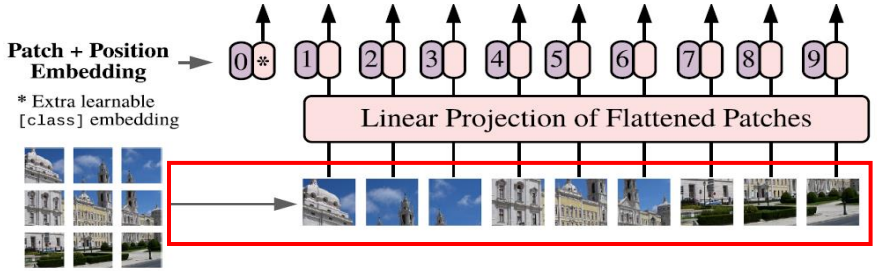
Q&A

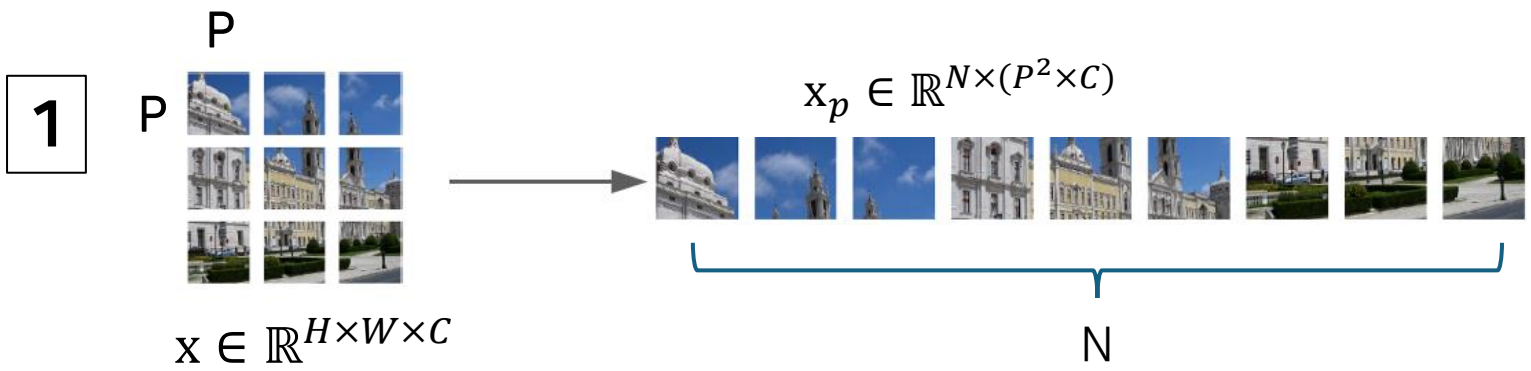
AIR Lab
서울과학기술대학교 인공지능 / 환경원격탐사 연구실

Appendix

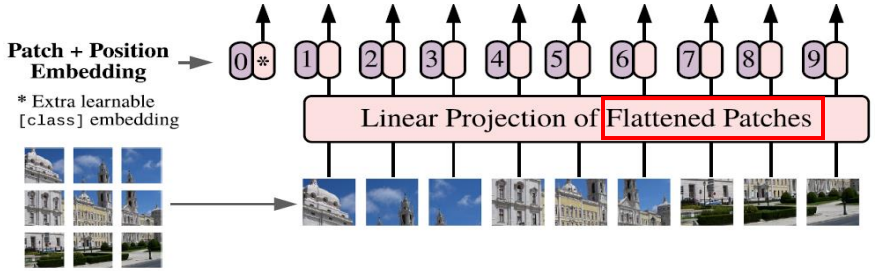
1. ViT 모델 구조

Step 1. 이미지 $\mathrm{x} \in \mathbb{R}^{H \times W \times C}$가 있을 때, 이미지를 $(P \times P)$ 크기의 패치 N개로 분할하여 패치 sequence $\mathrm{x}_p \in \mathbb{R}^{N \times (P^2 \times C)}$를 구축함

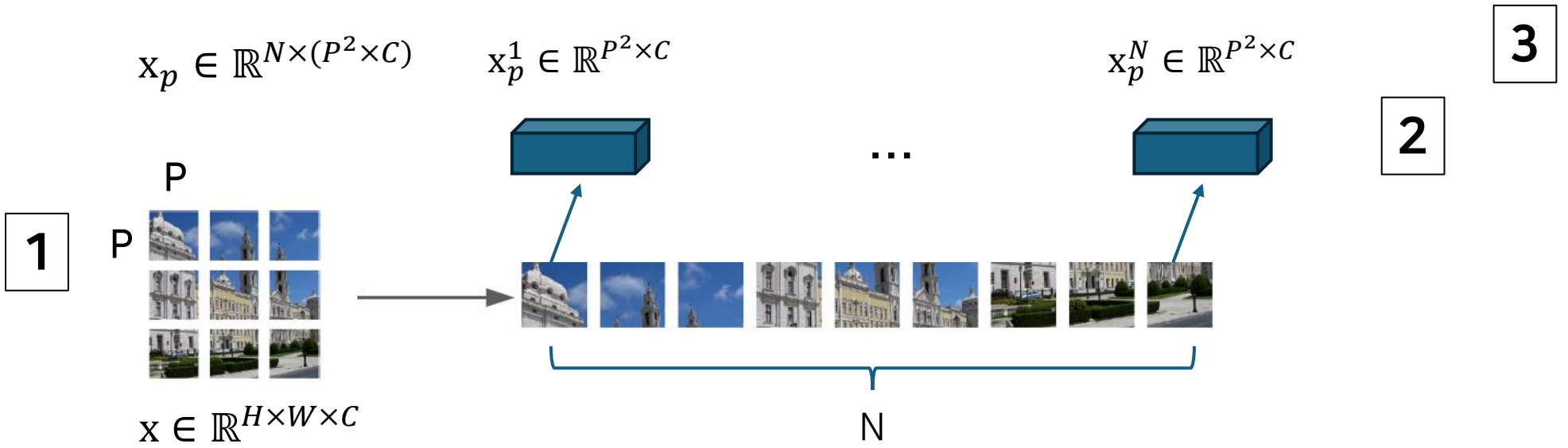




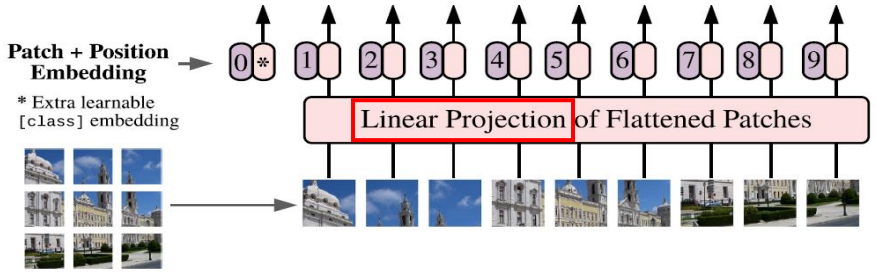
1. ViT 모델 구조



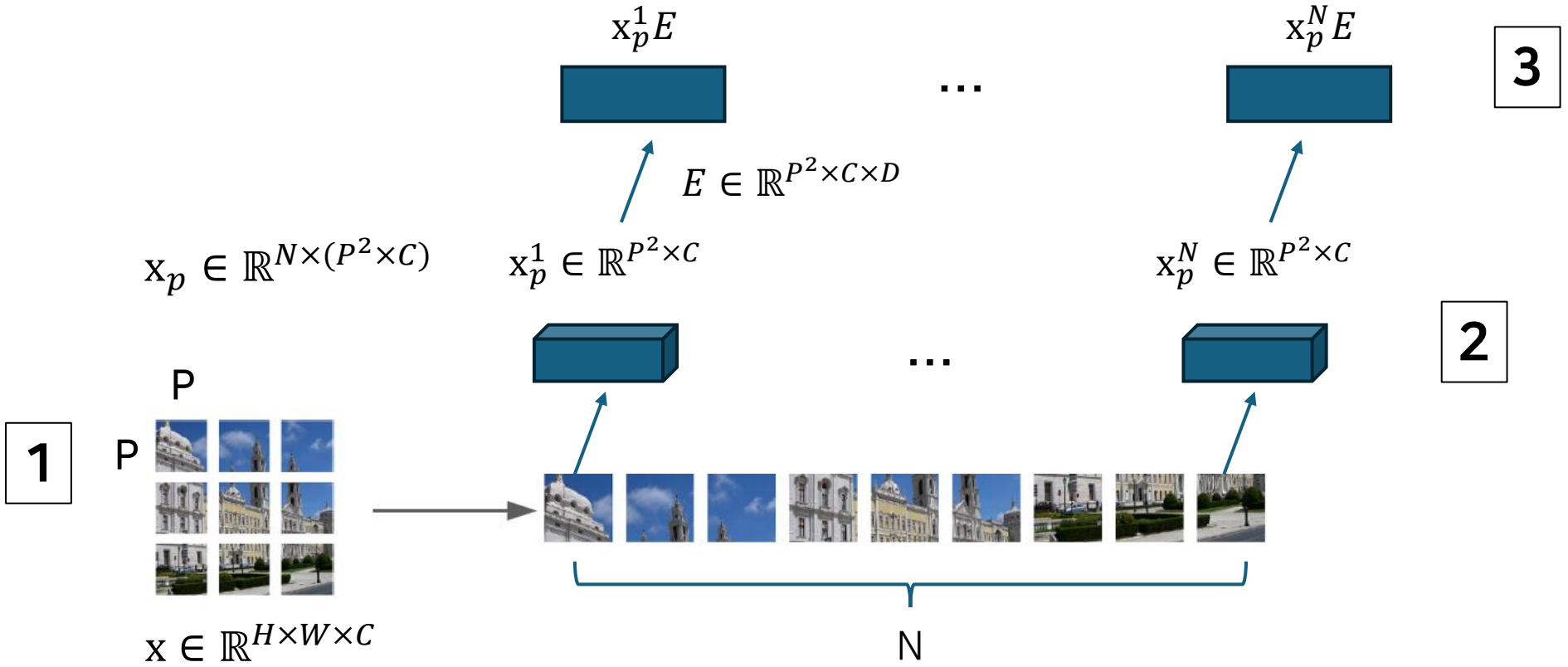
Step 2. Trainable linear projection을 통해 x_p의 각 패치를 flatten한 벡터를 D차원으로 변환한 후, 이를 패치 임베딩으로 사용



1. ViT 모델 구조



Step 2. Trainable linear projection을 통해 x_p의 각 패치를 flatten한 벡터를 D차원으로 변환한 후, 이를 패치 임베딩으로 사용



1. ViT 모델 구조

Step 2. Trainable linear projection을 통해 x_p의 각 패치를 flatten한 벡터를 D차원으로 변환한 후, 이를 패치 임베딩으로 사용