Transformer in Computer Vision

Open DMQA Seminar
2021.03.26

조한샘

고려대학교
KOREA UNIVERSITY

발표자 소개

- 조한샘

  Data Mining & Quality Analytics Lab

  석사과정 (2020.09~)

- 관심 연구 분야

  Generative Models

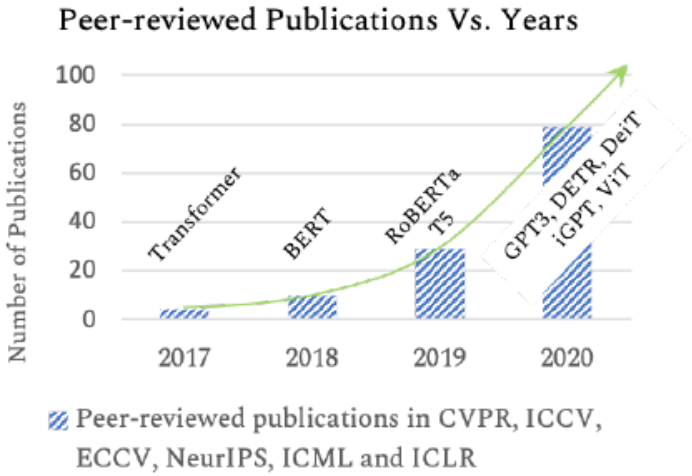
- E-mail: chosam95@korea.ac.kr

CONTENTs

Introduction

RNN to Transformer in NLP

- RNN은 오랜시간동안 NLP 분야에서 기본 모델로써 활용됨
- Transformer의 등장 이후 Transformer 중심으로 연구 진행



https://arxiv.org/pdf/2101.01169.pdf
https://www.youtube.com/watch?v=S27pHKBEp30

Introduction

Transformer in Computer Vision

- Non-local neural networks (Wang et al., 2018)
- Stand-alone self-attention in vision models (Ramachandran et al.,2019)
- Axial-DeepLab (Wang et al., 2020)

⋮

- Vision Transformer (Dosovitskiy et al., 2020)
- Data efficient image Transformer (Touvron et al., 2020)
- TransGAN (Jiang et al., 2021)

⋮

CNN에 self attention을 어떻게 적용할까?

Transformer 모델 자체를 이용해보자!

Transformer

Transformer and Self Attention

- Transformer: Attention만을 활용해 모델 구축
- Transformer의 핵심 아이디어 → Self Attention

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

NIPS (2017)

인용횟수: 18,767 (2021.03.23)

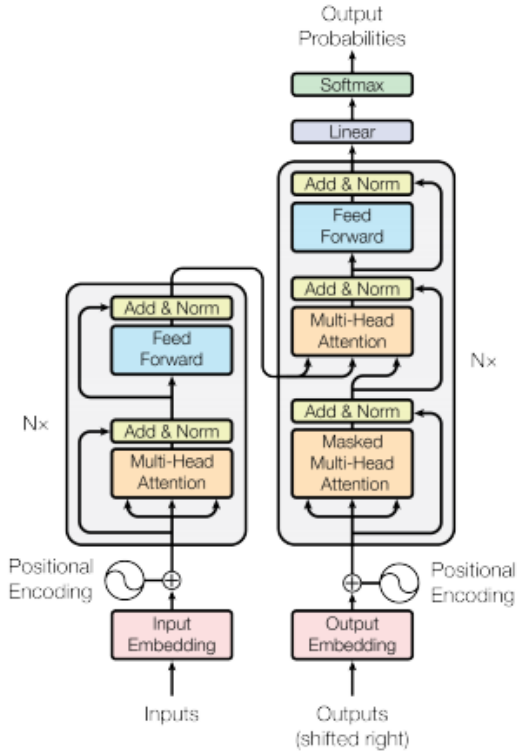
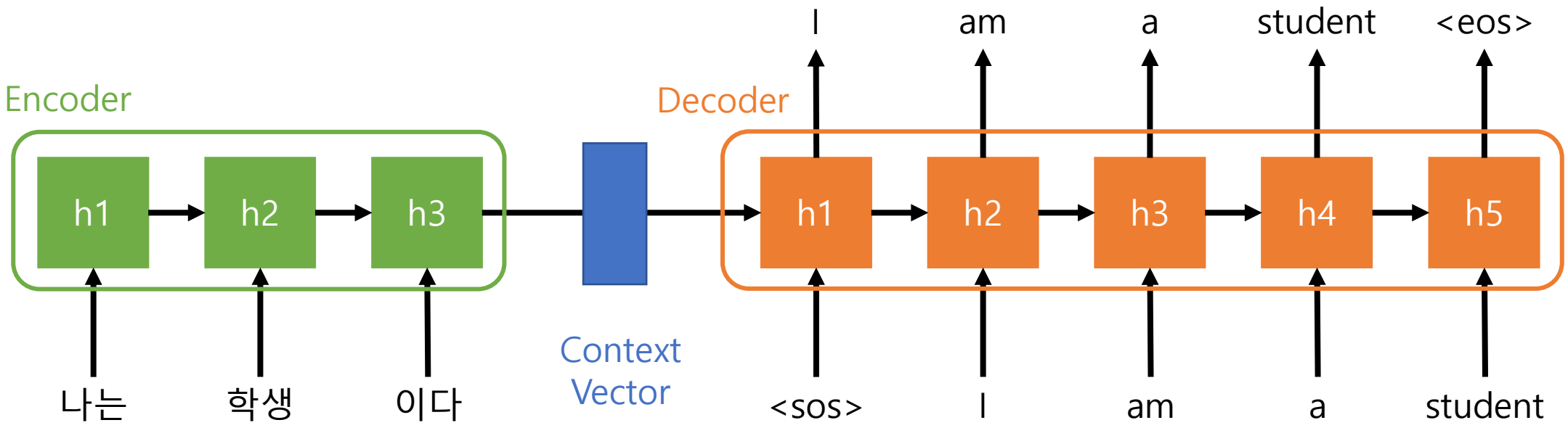


Figure 1: The Transformer - model architecture.

Transformer

Seq2seq

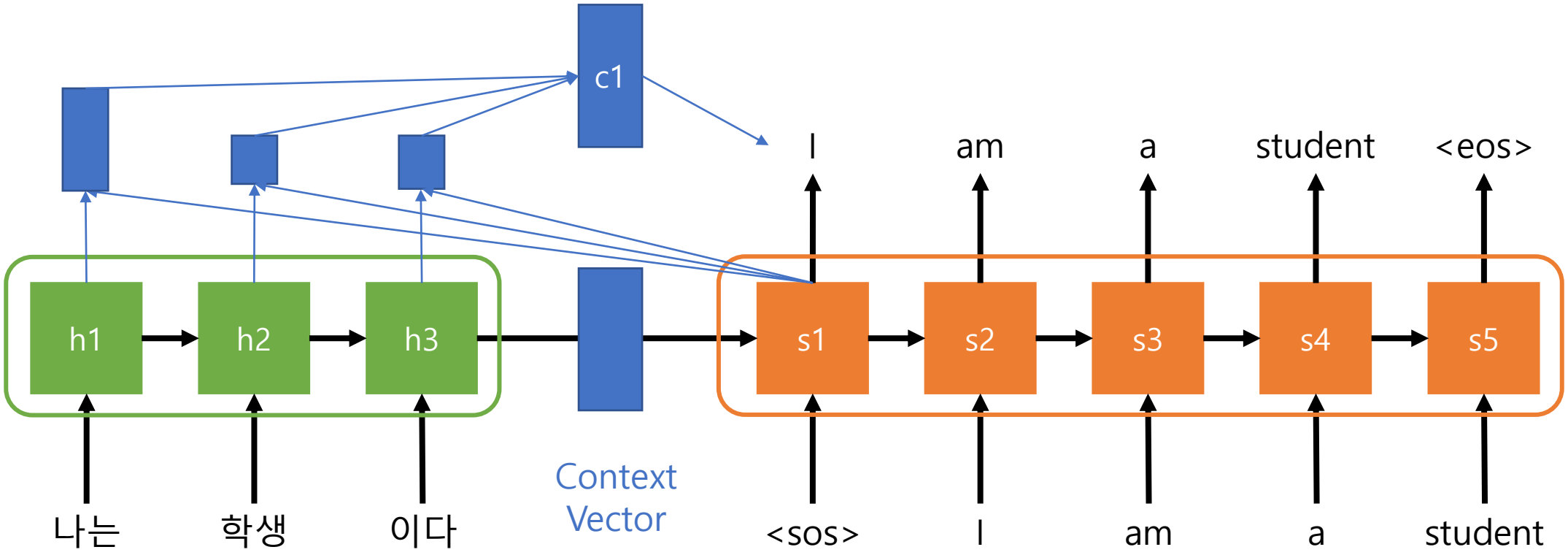
- Seq2seq: 문장을 입력으로 받아 문장을 출력하는 모델 / 기계번역에 주로 사용
- Context vector: Decoder에게 전달되는 입력 문장의 정보
- Context vector의 크기가 제한적이기 때문에 입력 문장의 모든 정보를 전하기 어렵다



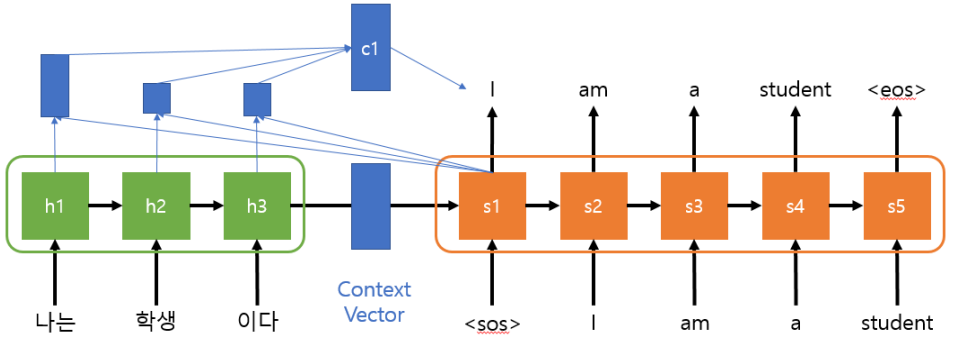
Transformer

Seq2seq with Attention

- Decoder가 특정 시점 단어를 출력할 때 encoder 정보 중 연관성이 있는 정보를 직접 선택

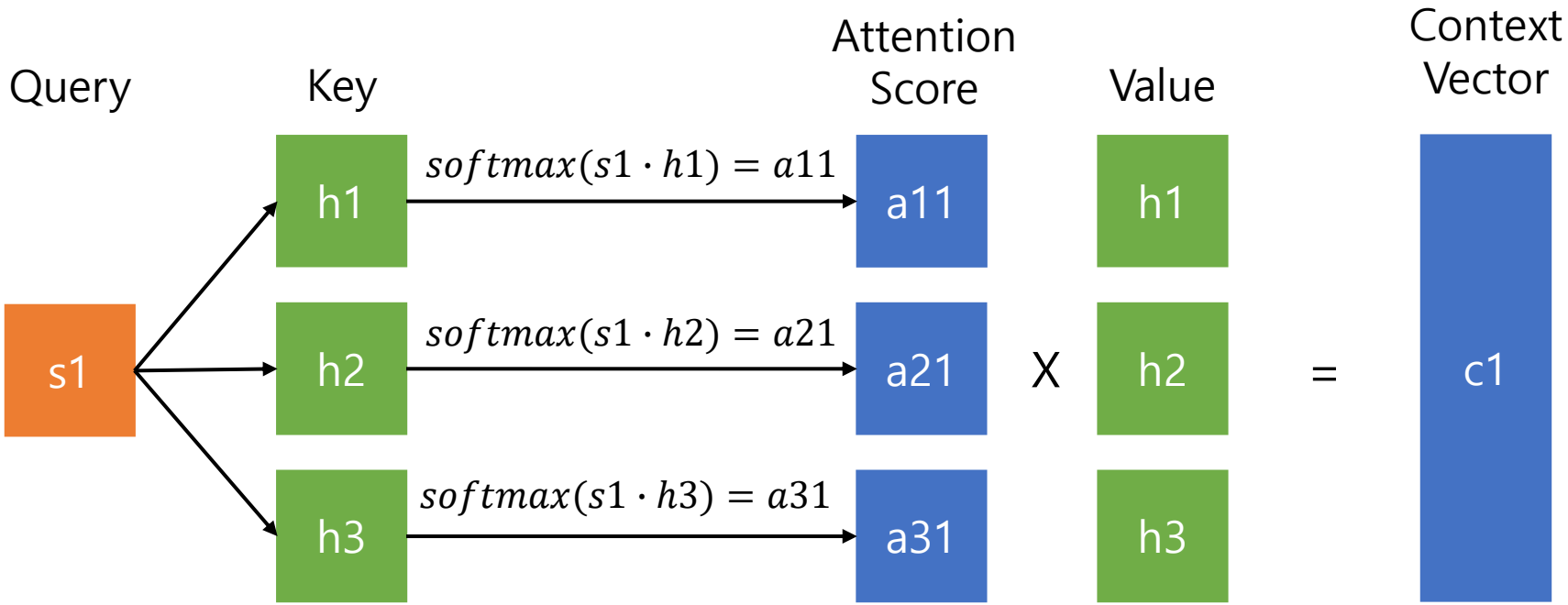


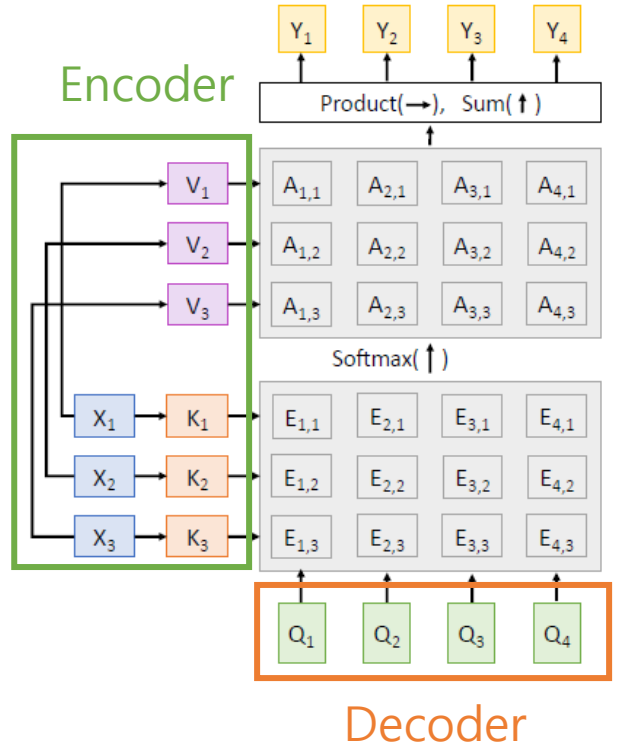
Transformer



Seq2seq with Attention

- Decoder가 특정 시점 단어를 출력할 때 encoder 정보 중 연관성이 있는 정보를 직접 선택





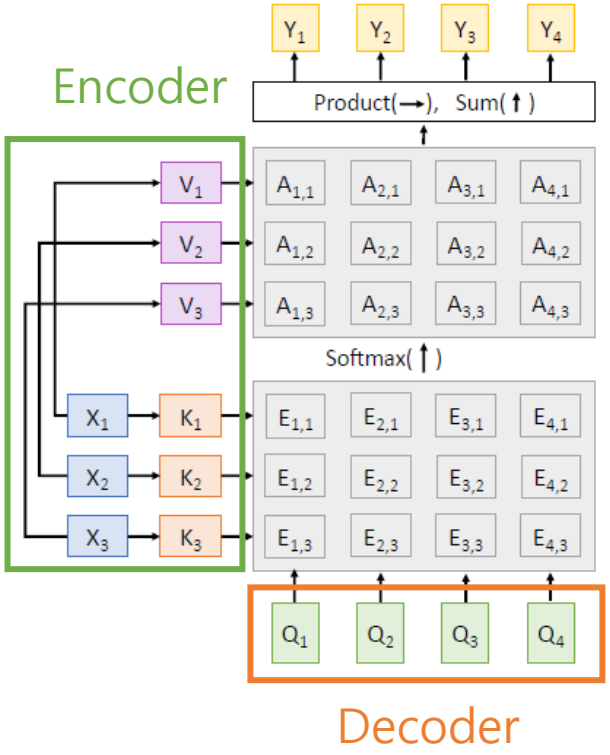
$$c1 = a11h1 + a21h2 + a31h3$$

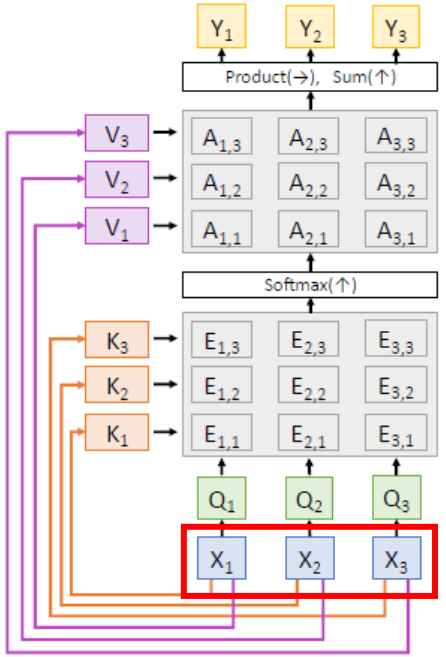
https://web.eecs.umich.edu/~justincj/slides/eecs498/498_FA2019_lecture13.pdf

Transformer

Attention vs Self Attention

- Attention (Decoder → Query / Encoder → Key, Value) / encoder, decoder 사이의 상관관계를 바탕으로 특징 추출
- Self attention (입력 데이터 → Query, Key, Value) / 데이터 내의 상관관계를 바탕으로 특징 추출



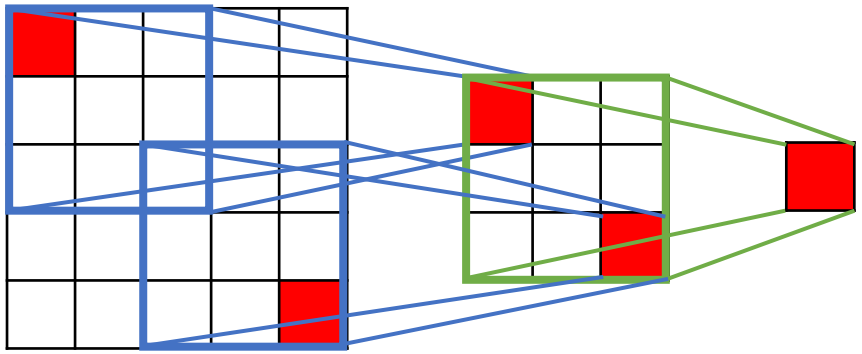


https://web.eecs.umich.edu/~justincj/slides/eecs498/498_FA2019_lecture13.pdf

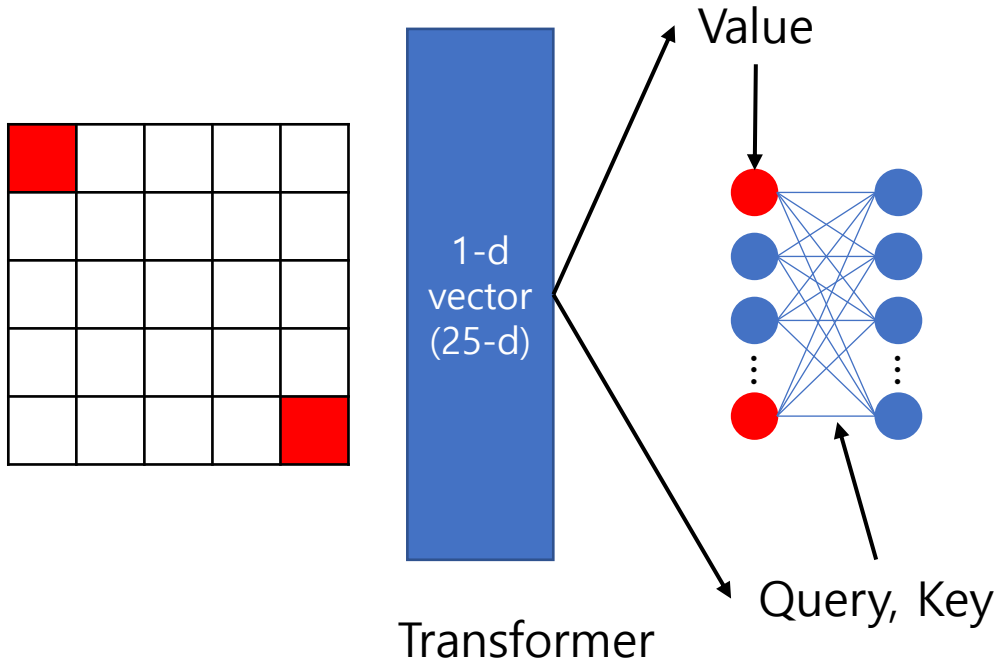
Transformer

Transformer vs CNN

- CNN: 이미지 전체의 정보를 통합하기 위해서는 몇 개의 layer 통과
- Transformer: 하나의 layer로 전체 이미지 정보 통합 가능



CNN

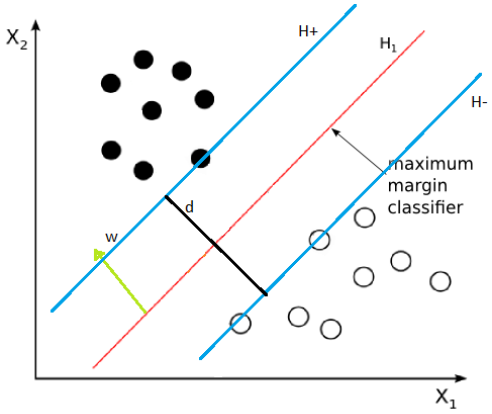


Transformer

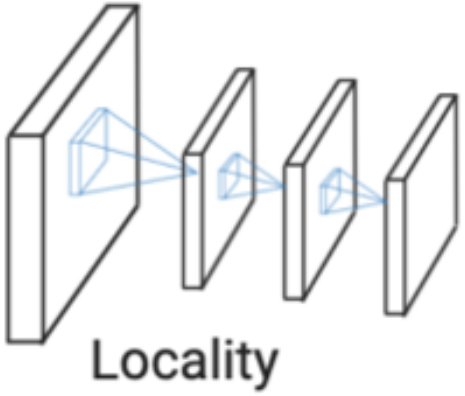
Transformer

Inductive bias

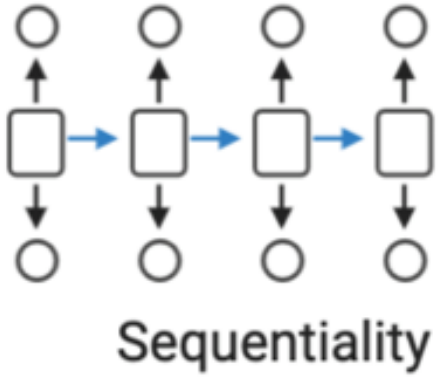
- Inductive bias: 새로운 데이터에 대해 좋은 성능을 내기 위해 모델에 사전적으로 주어지는 가정
- SVM: Margin 최대화 / CNN: 지역적인 정보 / RNN: 순차적인 정보



SVM



CNN



RNN

https://medium.com/@surajdesai1297/whats-under-the-hood-of-svm-41ec089e7406
https://sgfin.github.io/2020/06/22/Induction-Intro/

Transformer

Inductive bias

- Transformer

  1차원 벡터로 만든 후 self attention (2차원의 지역적인 정보 유지 X)

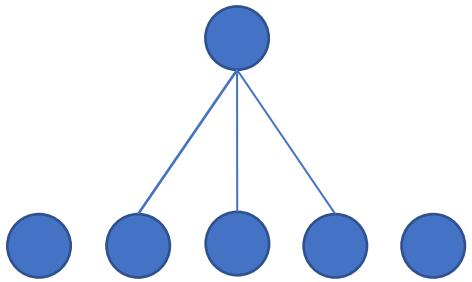
  Weight이 input에 따라 유동적으로 변함

- CNN

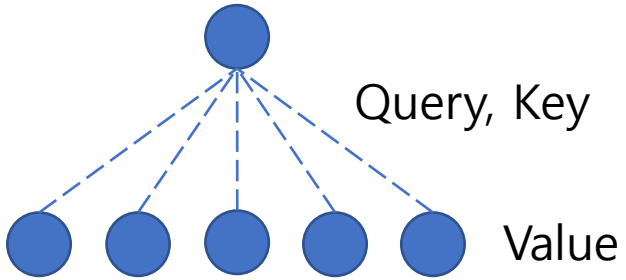
  2차원의 지역적인 특성 유지

  학습 후 weight이 고정

- Transformer: inductive bias ↓ , 모델의 자유도 ↑



CNN



Transformer

Transformer in Computer Vision

Vision Transformer (ViT)

- 103회 인용 (21.03.24 기준)
- Transformer 구조를 활용해 image classification을 수행
- SOTA의 CNN 기반 모델과 비슷한 성능

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

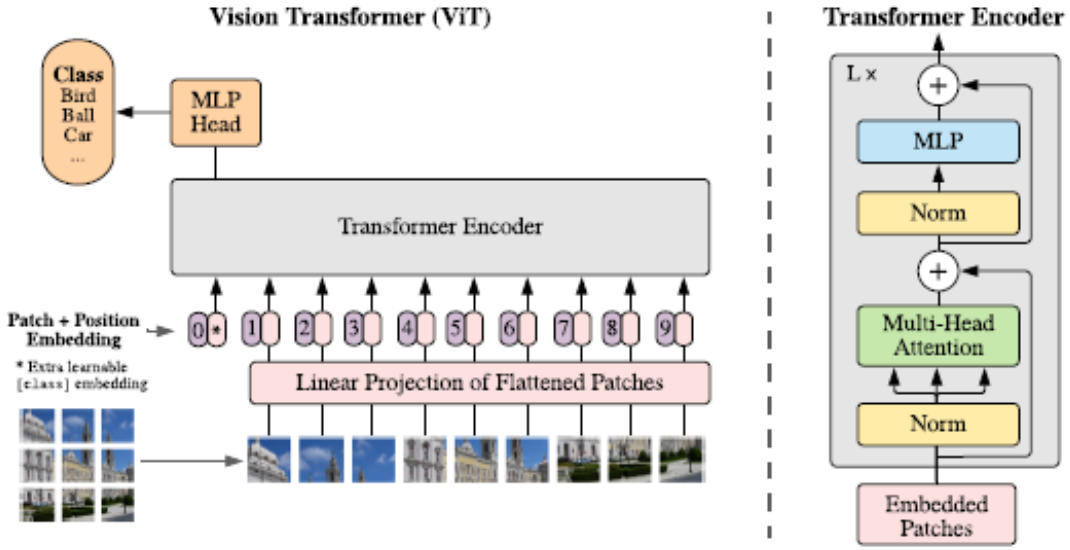
**Alexey Dosovitskiy*,†, Lucas Beyer*, Alexander Kolesnikov*, Dirk Weissenborn*,
Xiaohua Zhai*, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer,
Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby*,†**
*equal technical contribution, †equal advising
Google Research, Brain Team
{adosovitskiy, 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]

Transformer in Computer Vision

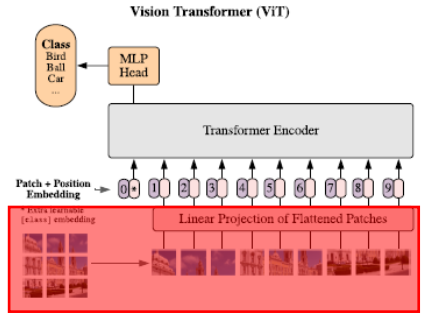
Vision Transformer (ViT)



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.

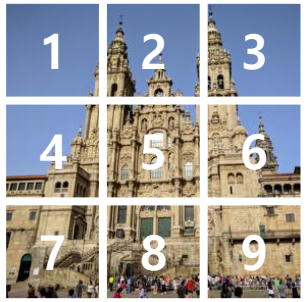
Transformer in Computer Vision



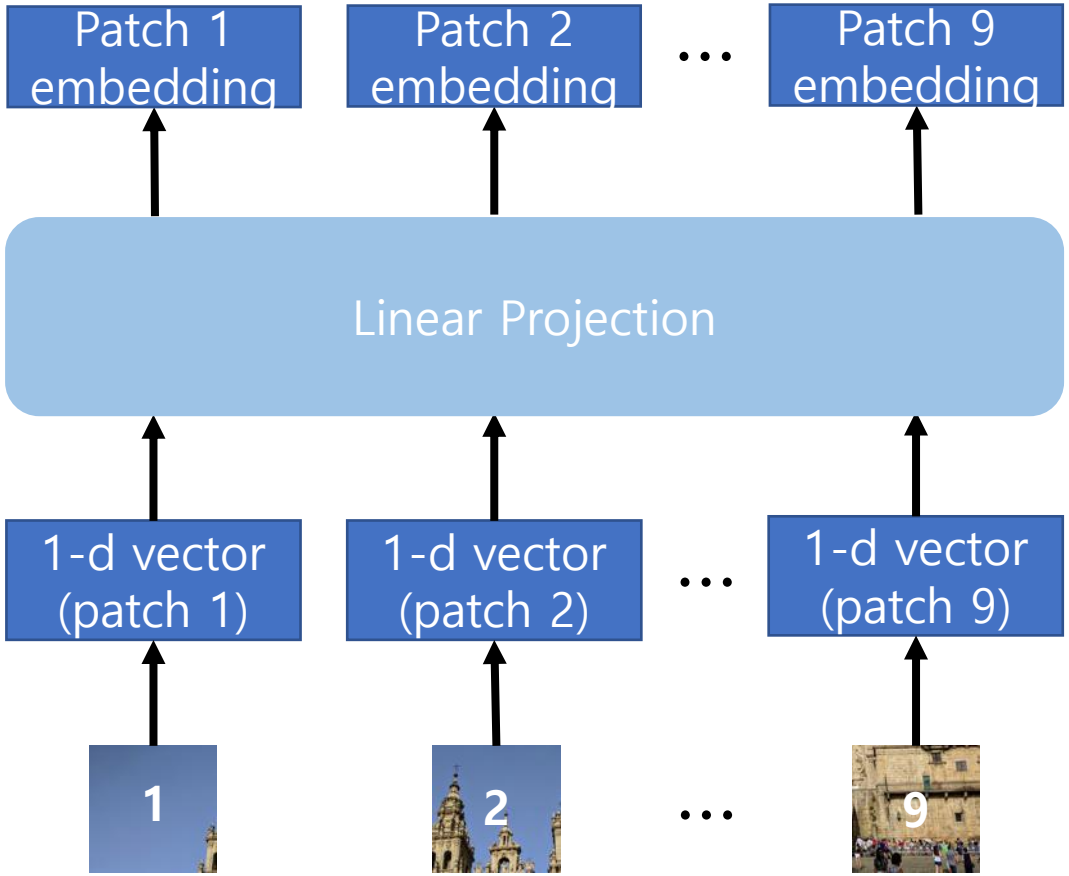
Vision Transformer example (ViT-Base/16)

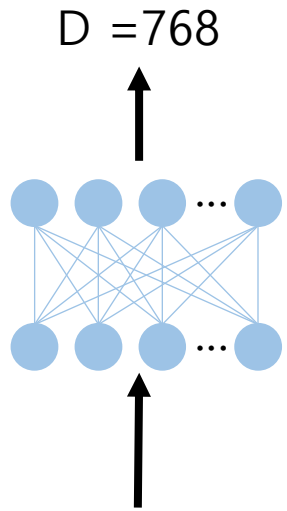
Input Image

48X48

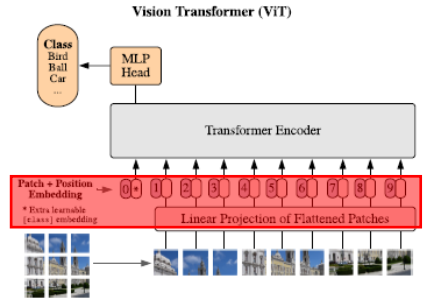


(16X16)*9





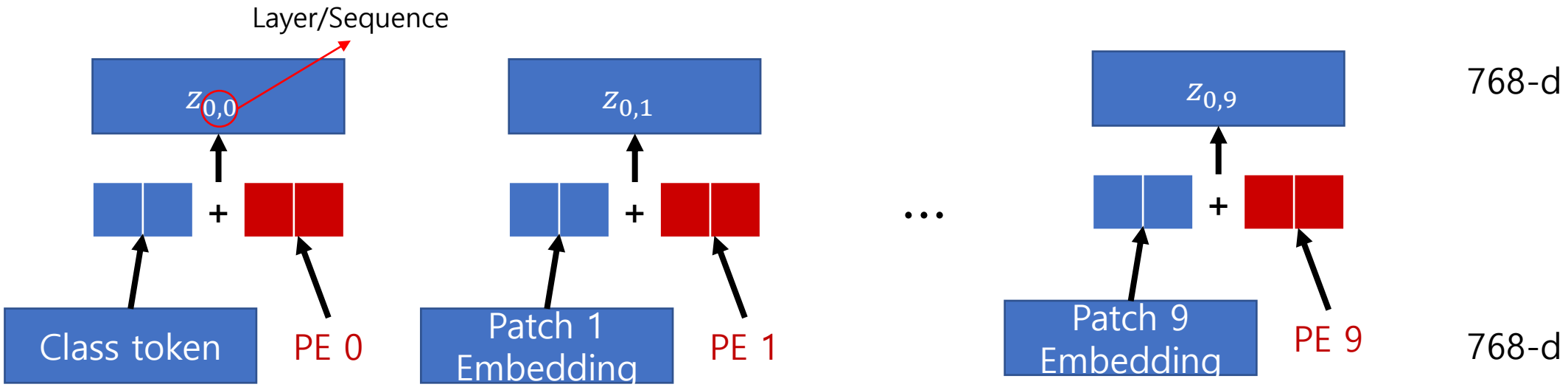
Transformer in Computer Vision



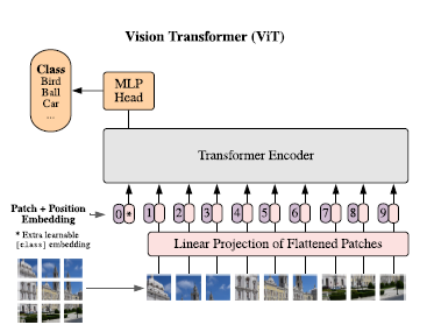
Vision Transformer example (ViT-Base/16)

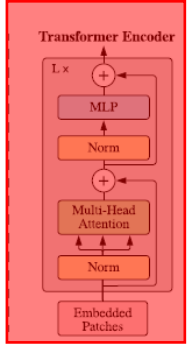
- **Classification token**: classification을 위해 사용되는 token (BERT [CLS] token)
- **Position embedding**: patch의 위치 정보
- (Classification token, Patch embedding) + Positional embedding = Transformer encoder 입력

학습을 통해 결정



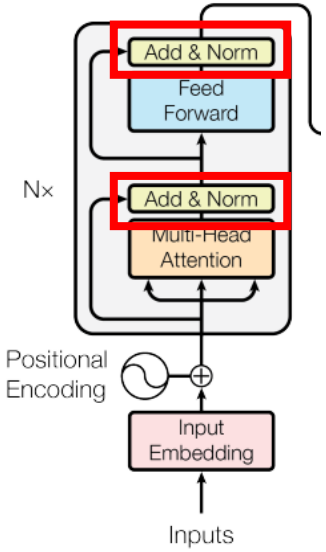
Transformer in Computer Vision



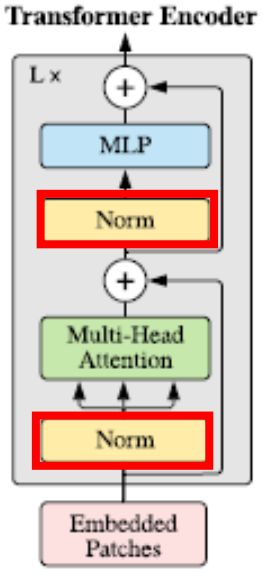


Vision Transformer example (ViT-Base/16)

- "Vanilla" Transformer encoder vs "ViT" Transformer encoder
- Layer normalization의 위치가 Transformer 학습에 중요한 역할 (Wang et al., 2019)
- ViT는 수정된 Transformer encoder 적용



"Vanilla" Transformer

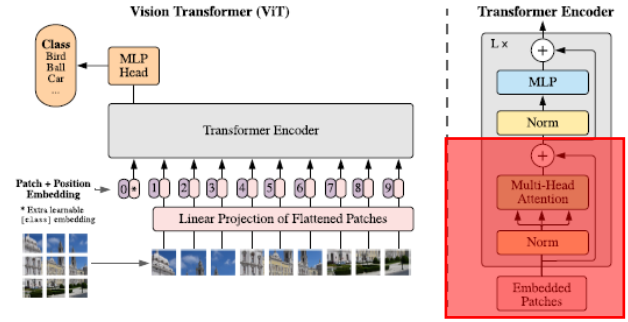


"ViT" Transformer

Learning Deep Transformer Models for Machine Translation

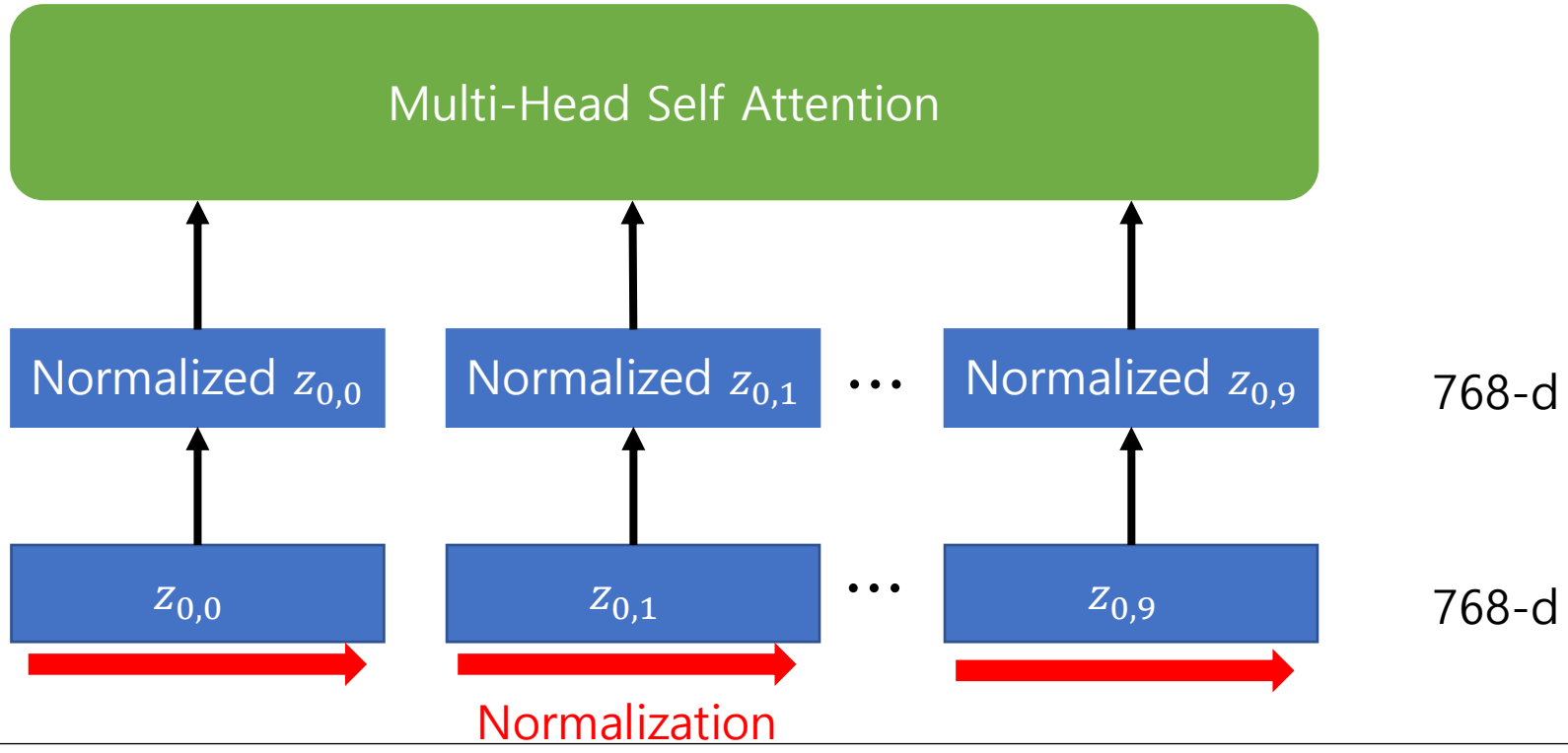
Qiang Wang[1], Bei Li[1], Tong Xiao[1,2]*, Jingbo Zhu[1,2], Changliang Li[3], Derek F. Wong[4], Lidia S. Chao[4]
[1]NLP Lab, Northeastern University, Shenyang, China
[2]NiuTrans Co., Ltd., Shenyang, China
[3]Kingsoft AI Lab, Beijing, China
[4]NLP²CT Lab, University of Macau, Macau, China
wangqiangneu@gmail.com, libei_neu@outlook.com,
{xiaotong,zhujingbo}@mail.neu.edu.com,
lichangliang@kingsoft.com, {derekfw,lidiasc}@um.edu.mo

Transformer in Computer Vision



Vision Transformer example (ViT-Base/16)

- Transformer encoder: Layer normalization

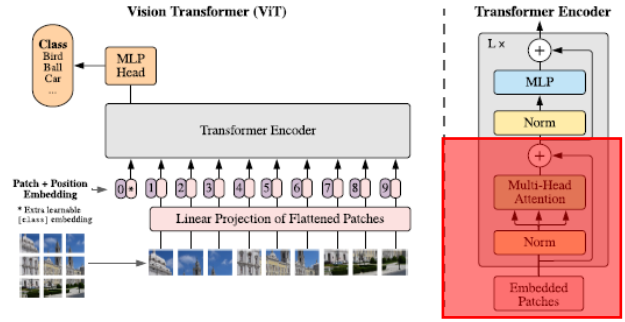


	X_1	X_2	X_3	X_4
1				
2		Data		
3				

Layer Normalization

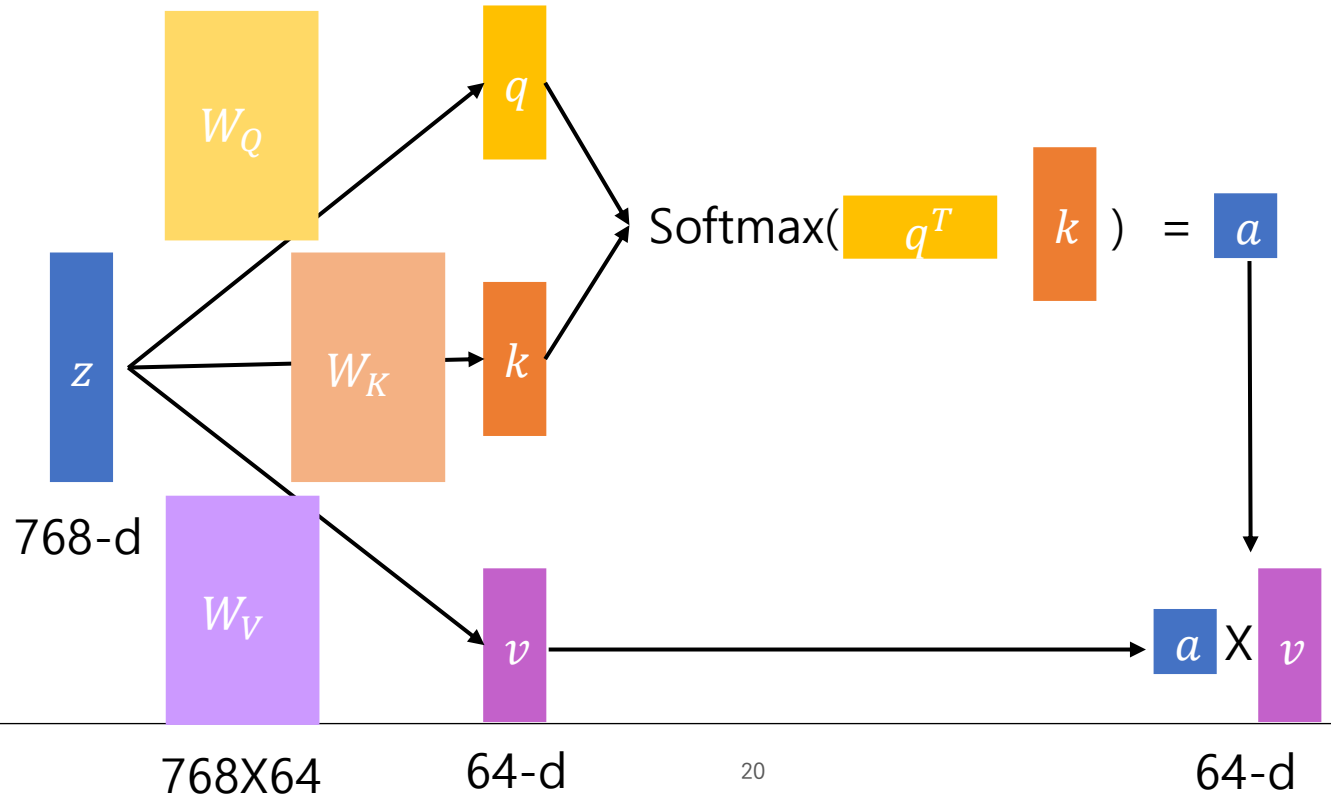
Batch Normalization

Transformer in Computer Vision

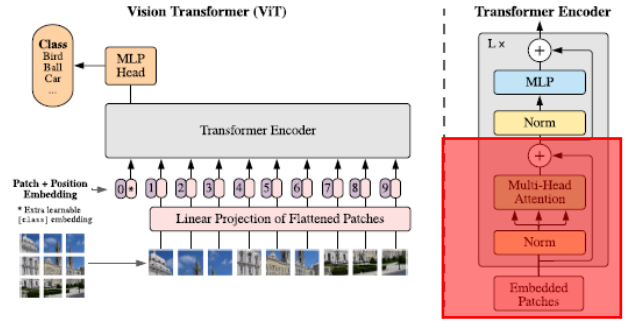


Vision Transformer example (ViT-Base/16)

- Transformer encoder: Self attention
- Encoder의 입력(z) → query, key, value 벡터 / W matrix: 학습되는 파라미터



Transformer in Computer Vision

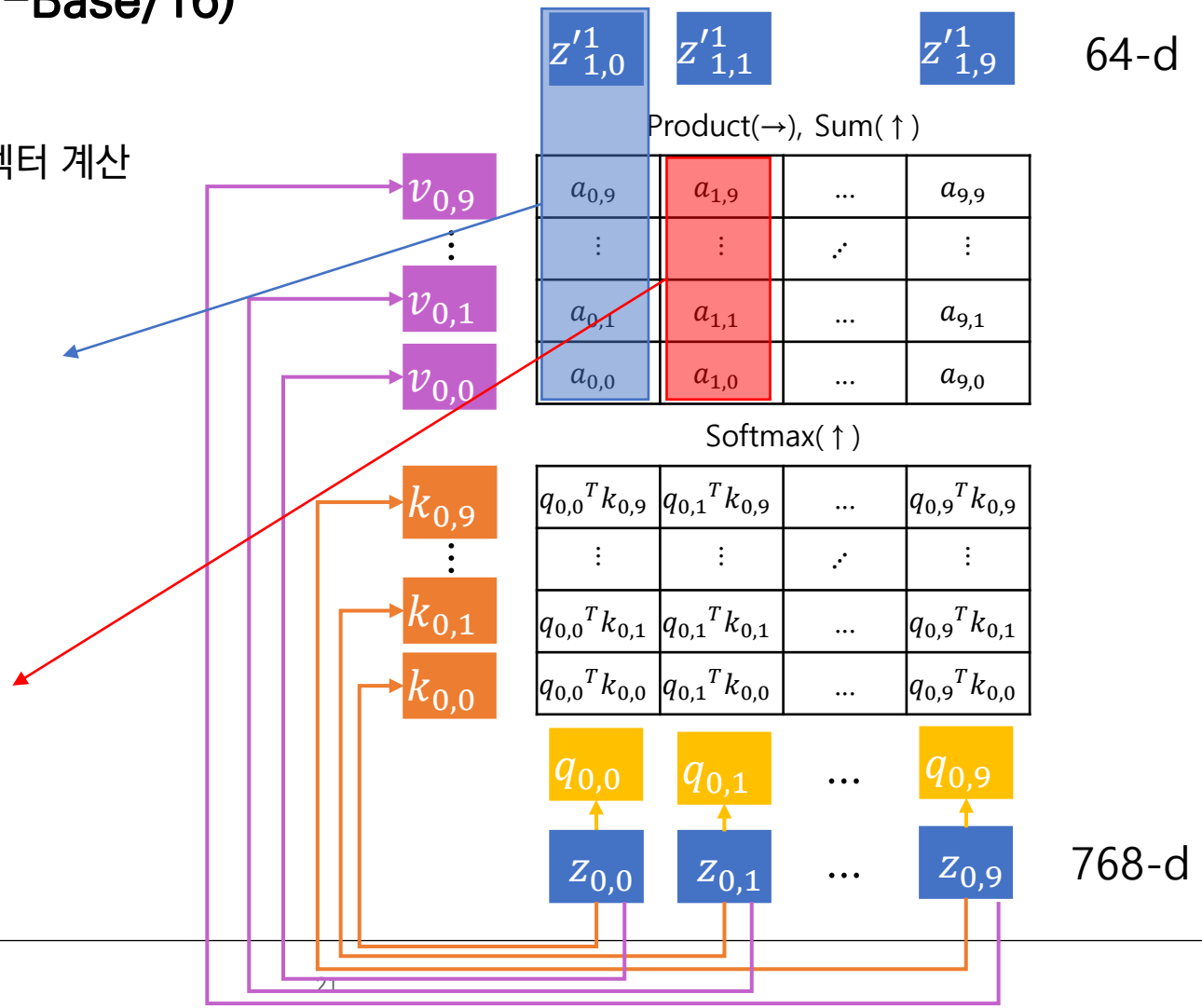


Vision Transformer example (ViT-Base/16)

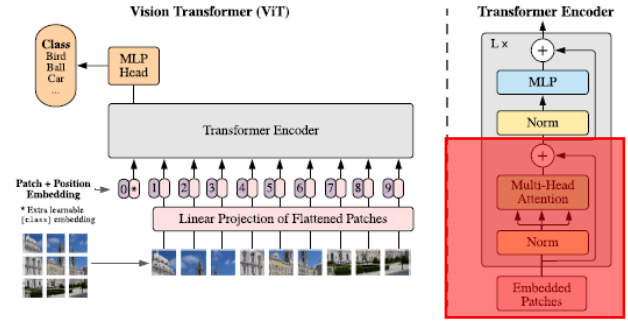
- Transformer encoder: Self attention
- Encoder의 입력(z) → query, key, value 벡터 계산

$$z'^{1}_{1,0} = a_{0,0}v_{0,0} + a_{0,1}v_{0,1} + \cdots + a_{0,9}v_{0,9}$$

1번 patch 입장에서 다른
embedding vector들과의 유사도

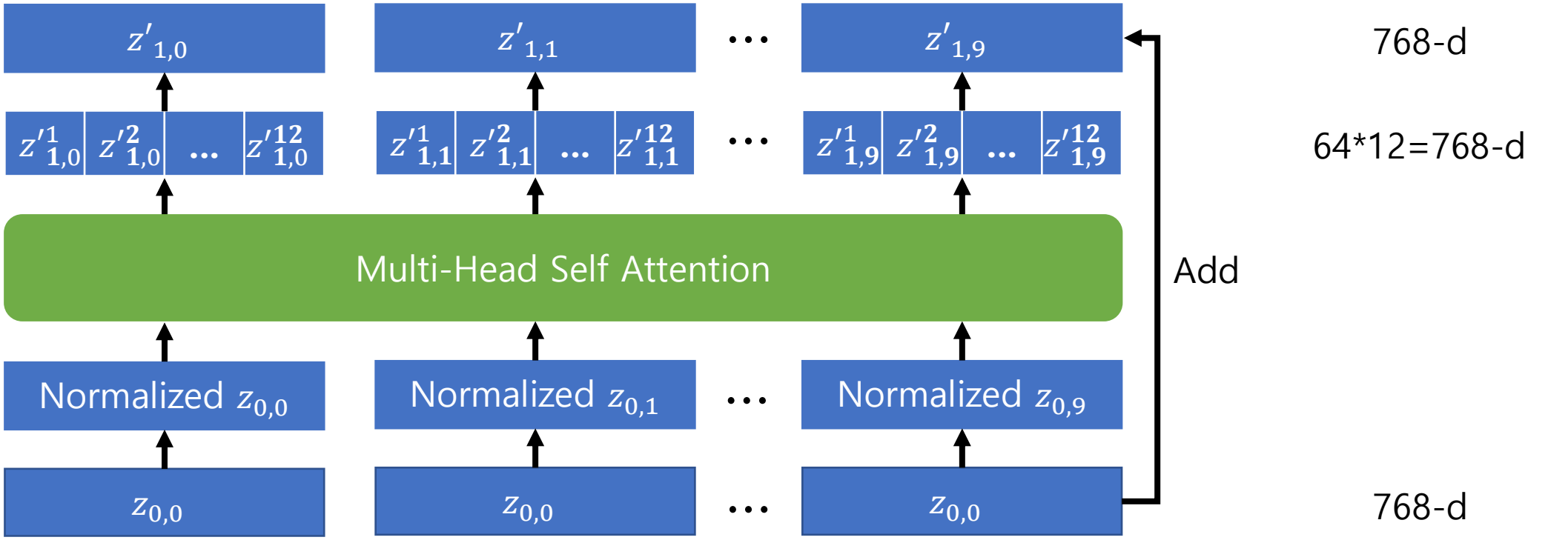


Transformer in Computer Vision

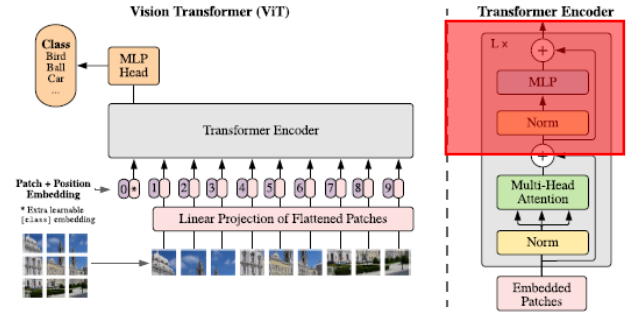


Vision Transformer example (ViT-Base/16)

- Transformer encoder: Multi-head self attention
- Self attention 12번 수행

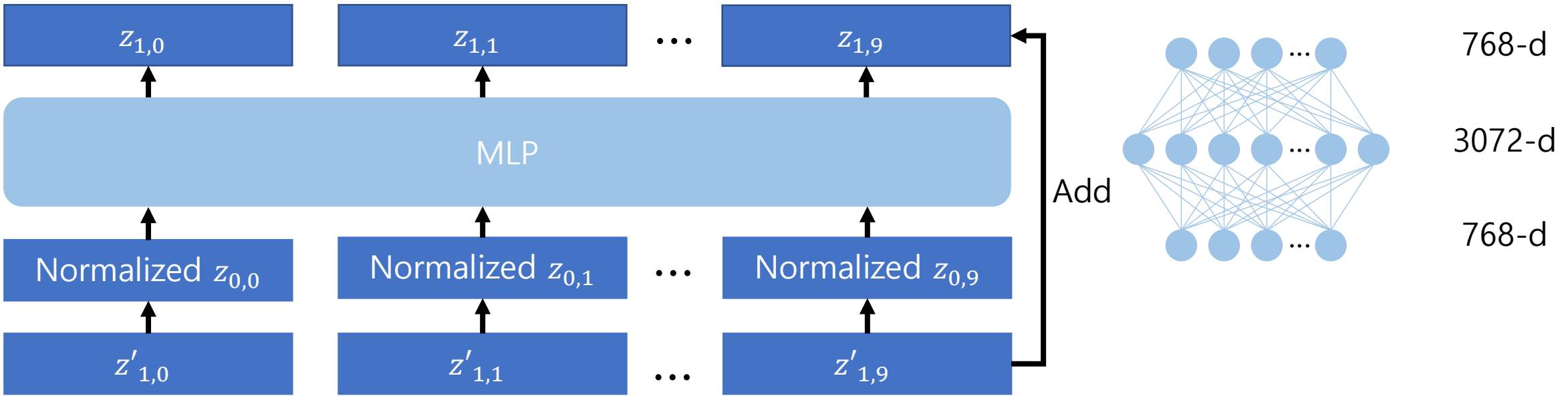


Transformer in Computer Vision

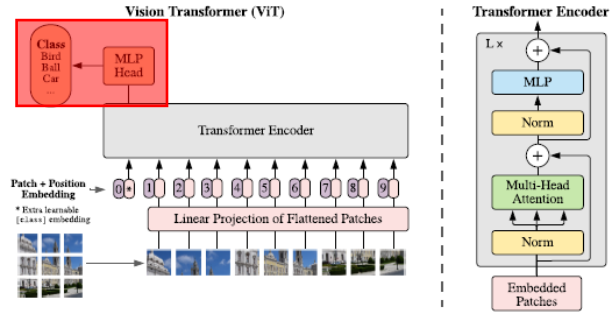


Vision Transformer example (ViT-Base/16)

- Transformer encoder: MLP

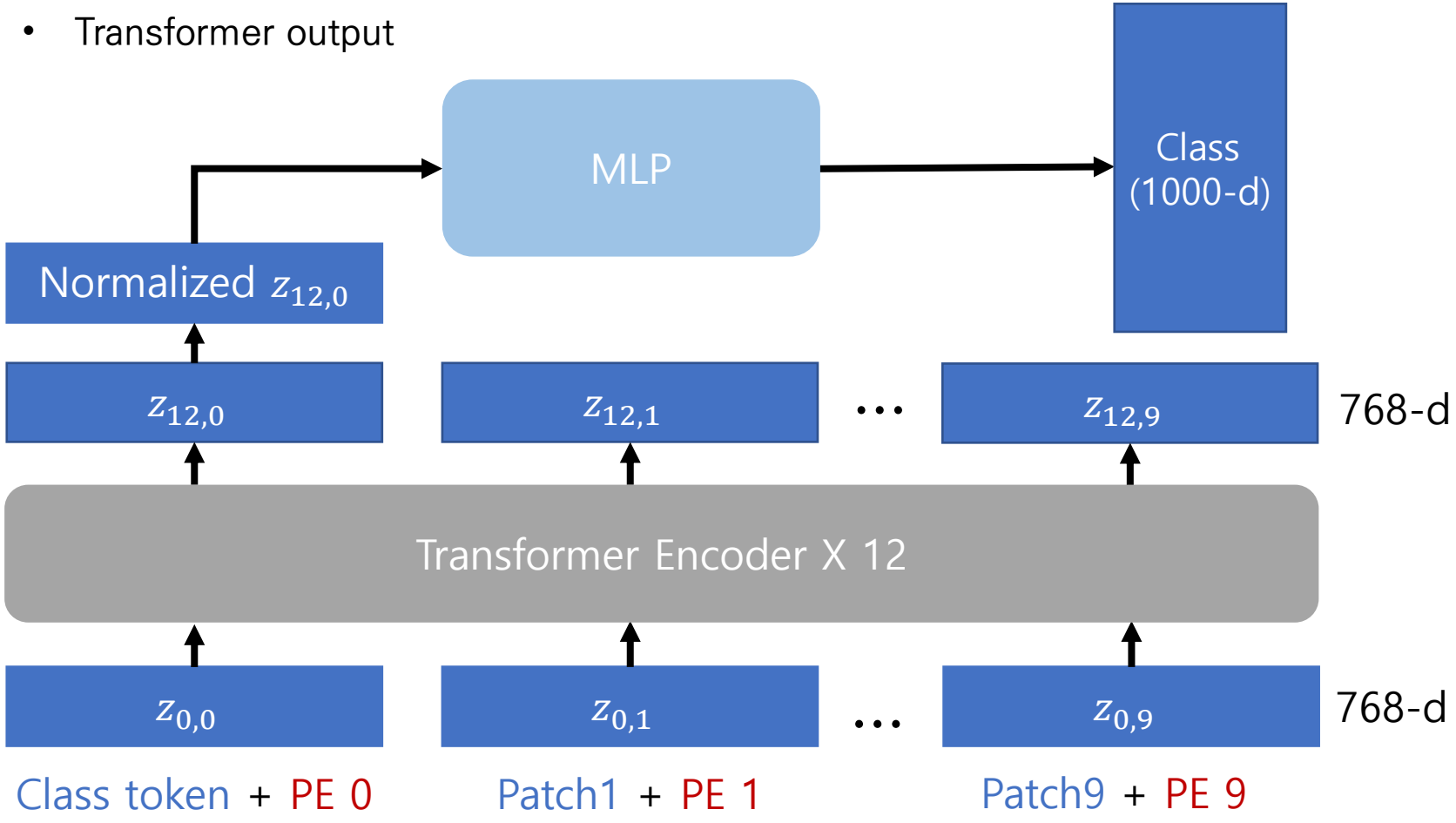


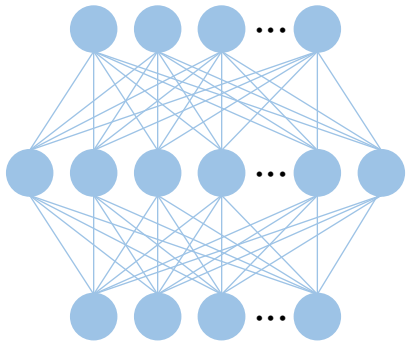
Transformer in Computer Vision

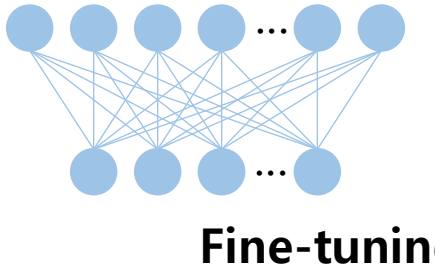


Vision Transformer example (ViT-Base/16)

- Transformer output







Transformer in Computer Vision

Vision Transformer Experiments

- Transfer learning 성능 비교
- Pre-training 이미지 resolution(224*224) → Fine-tuning 이미지 resolution(384*384) (Touvron et al., 2019)
- Big Transfer보다 사용하는 자원↓/성능↑ → 효율적인 사전 학습 가능

Pre-training dataset
(Pre-training model)

Fine-tuning dataset	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

Fixing the train-test resolution discrepancy

Hugo Touvron, Andrea Vedaldi, Matthijs Douze, Hervé Jégou

Facebook AI Research

Abstract

Data-augmentation is key to the training of neural networks for image classification. This paper first shows that existing augmentations induce a significant discrepancy between the size of the objects seen by the classifier at train and test time: in fact, a lower train resolution improves the classification at test time!

We then propose a simple strategy to optimize the classifier performance, that employs different train and test resolutions. It relies on a computationally cheap fine-tuning of the network at the test resolution. This enables training strong classifiers using small training images, and therefore significantly reduce the training time. For instance, we obtain 77.1% top-1 accuracy on ImageNet with a ResNet-50 trained on 128×128 images, and 79.8% with one trained at 224×224.

A ResNeXt-101 32x48d pre-trained with weak supervision on 940 million 224×224 images and further optimized with our technique for test resolution 320×320 achieves 86.4% top-1 accuracy (top-5: 98.0%). To the best of our knowledge this is the highest ImageNet single-crop accuracy to date.

Transformer in Computer Vision

Vision Transformer Experiments

- Transfer learning 성능 비교
- Pre-training 이미지 resolution(224*224) → Fine-tuning 이미지 resolution(384*384) (Touvron et al., 2019)
- Big Transfer보다 사용하는 자원↓/성능↑ → 효율적인 사전 학습 가능

Pre-training dataset
(Pre-training model)

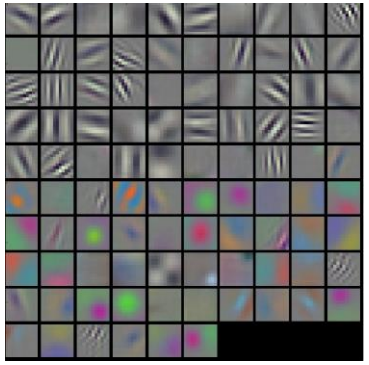
Fine-tuning dataset	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21K (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	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

Dataset	Classes	Images
ImageNet	1k	1.3M
ImageNet-21k	21k	14M
JFT	18k	303M

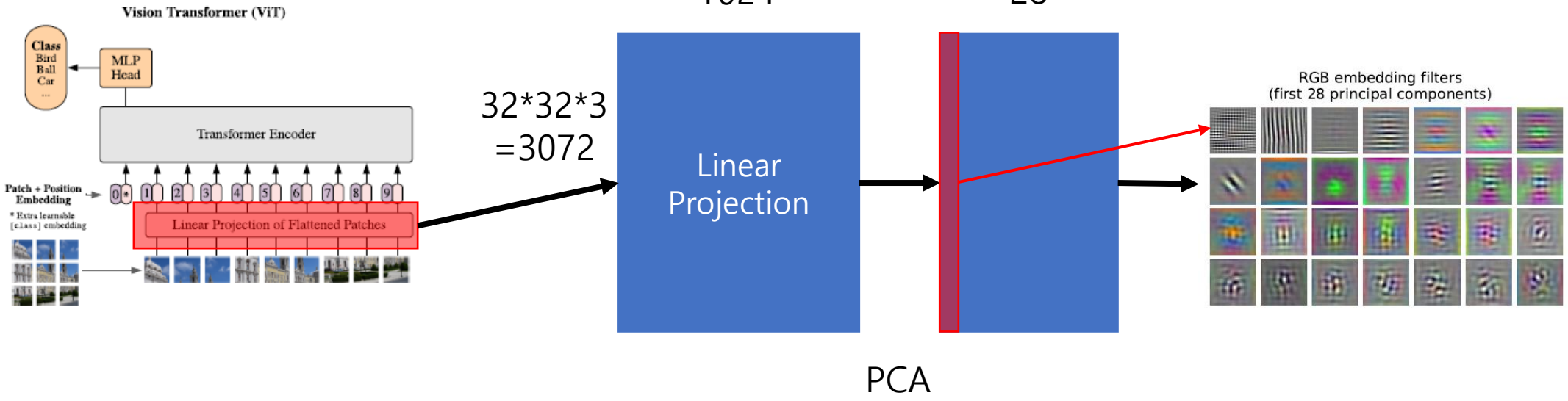
Transformer in Computer Vision

Vision Transformer Experiments

- Linear projection 시각화 (ViT-L/32)
- CNN의 convolution filter를 시각화한 것과 유사



CNN convolution filter



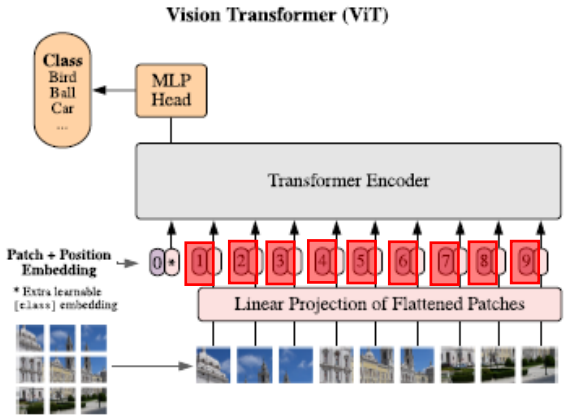
ViT-L/32
Linear Projection

https://cs231n.github.io/understanding-cnn/

Transformer in Computer Vision

Vision Transformer Experiments

- Position embedding 시각화 (ViT-L/32)
- 이미지내 가까운 patch의 position embedding과 유사도 ↑

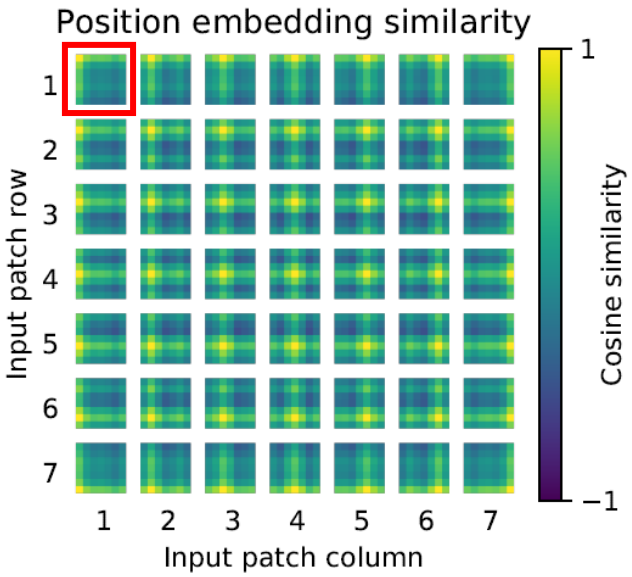


Input image (224X224)

1024-d

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31	32	33	34	35
36	37	38	39	40	41	42
43	44	45	46	47	48	49

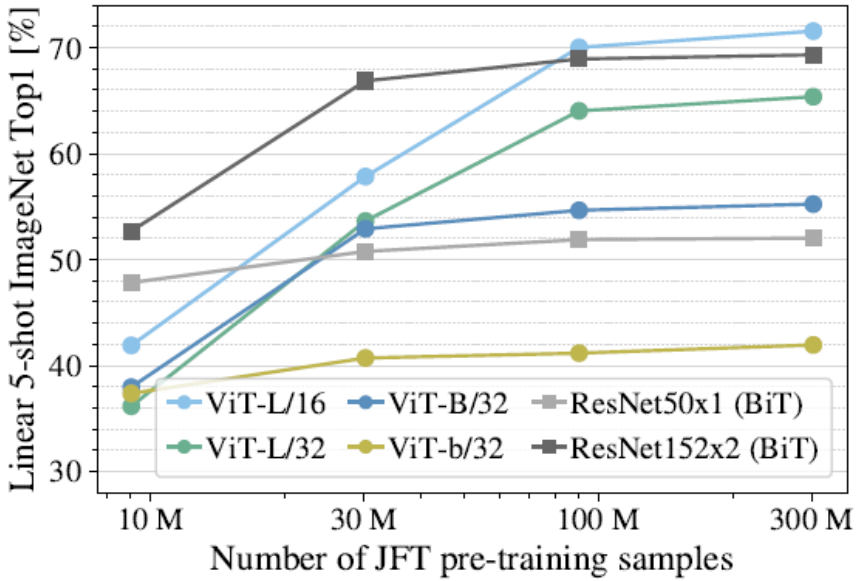
1번 PE과 나머지 PE 유사도 시각화



Transformer in Computer Vision

Vision Transformer Experiments

- 학습데이터가 충분하지 않을 경우 CNN 모델보다 성능 감소
- CNN에 비해 inductive bias ↓ / inductive bias까지 학습하기 위해 많은 양의 데이터 필요



Dataset	Classes	Images
ImageNet	1k	1.3M
ImageNet-21k	21k	14M
JFT	18k	303M

Transformer in Computer Vision

Data efficient image Transformer (DeiT)

- 21회 인용 (21.03.24 기준)
- 많은 데이터가 필요한 ViT 한계 극복

  → Knowledge Distillation & Data Augmentation

Training data-efficient image transformers & distillation through attention

Hugo Touvron*,† Matthieu Cord† Matthijs Douze*
Francisco Massa* Alexandre Sablayrolles* Hervé Jégou*

*Facebook AI †Sorbonne University

Abstract

Recently, neural networks purely based on attention were shown to address image understanding tasks such as image classification. These high-performing vision transformers are pre-trained with hundreds of millions of images using a large infrastructure, thereby limiting their adoption.

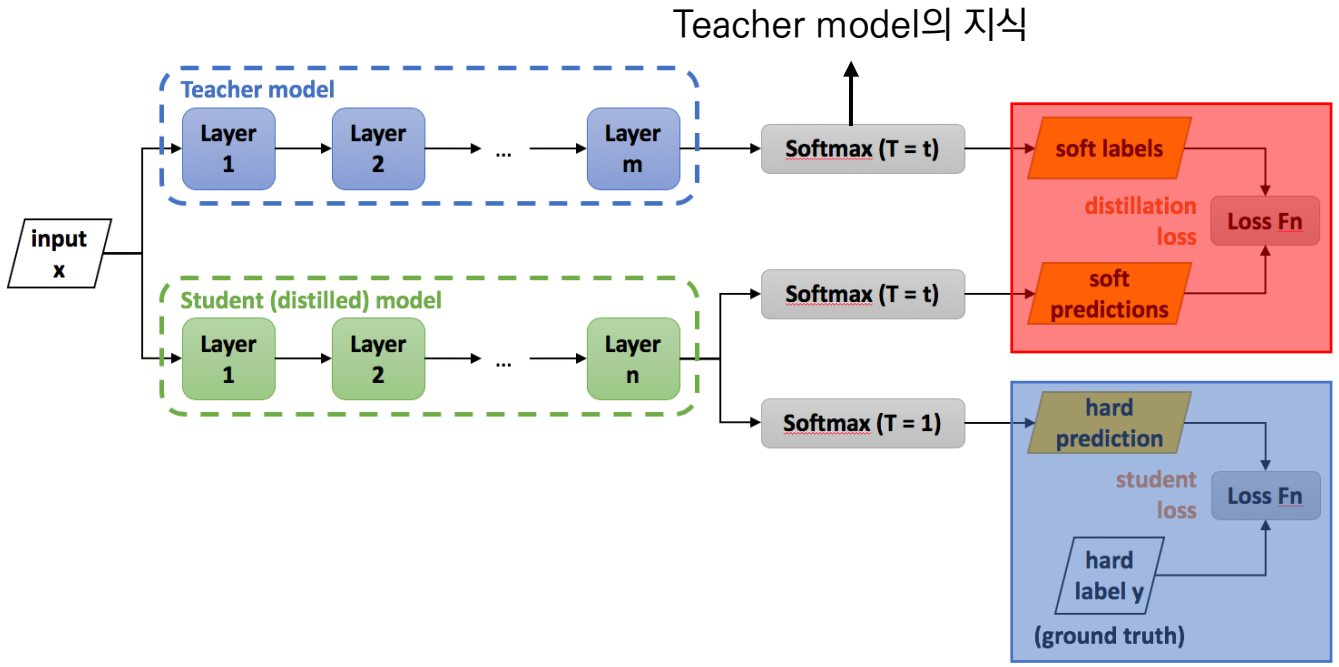
In this work, we produce competitive convolution-free transformers by training on Imagenet only. We train them on a single computer in less than 3 days. Our reference vision transformer (86M parameters) achieves top-1 accuracy of 83.1% (single-crop) on ImageNet with no external data.

More importantly, we introduce a teacher-student strategy specific to transformers. It relies on a distillation token ensuring that the student learns from the teacher through attention. We show the interest of this token-based distillation, especially when using a convnet as a teacher. This leads us to report results competitive with convnets for both Imagenet (where we obtain up to 85.2% accuracy) and when transferring to other tasks. We share our code and models.

Transformer in Computer Vision

Knowledge distillation

- 잘 학습된 모델의 지식을 전이 받아 좋은 성능을 내는 작은 모델 구축
- Teacher model: 특정 task와 data에 대해 잘 학습된 모델 (파라미터 ↑)
- Student model: Teacher model에게 지식을 전이 받는 모델 (파라미터 ↓)



$$\lambda\tau^2 KL\left(\psi\left(\frac{Z_s}{\tau}\right),\psi\left(\frac{Z_t}{\tau}\right)\right)$$

Distillation loss

$$+$$

$$(1-\lambda)L_{CE}(\psi(Z_s),y)$$

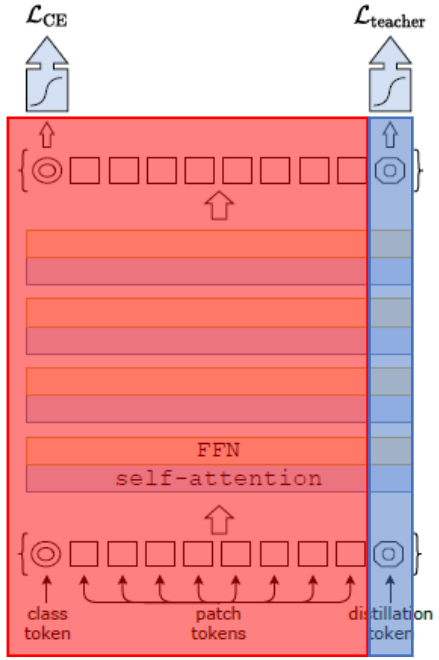
Classification loss

https://intellabs.github.io/distiller/knowledge_distillation.html

Transformer in Computer Vision

Data efficient image Transformer (DeiT)

- Vision Transformer와 동일한 구조
- Knowledge Distillation을 위한 Distillation token 추가 (Teacher model: RegNet)
- Classification token과 동일하게 학습을 통해 값을 결정



ViT

distillation token

Methods	ViT-B [15]	DeiT-B
Epochs	300	300
Batch size	4096	1024
Optimizer	AdamW	AdamW
learning rate	0.003	$0.0005 \times \frac{\text{batchsize}}{512}$
Learning rate decay	cosine	cosine
Weight decay	0.3	0.05
Warmup epochs	3.4	5
Label smoothing ε	✗	0.1
Dropout	0.1	✗
Stoch. Depth	✗	0.1
Repeated Aug	✗	✓
Gradient Clip.	✓	✗
Rand Augment	✗	9/0.5
Mixup prob.	✗	0.8
Cutmix prob.	✗	1.0
Erasing prob.	✗	0.25

Data Augmentation

Table 9: Ingredients and hyper-parameters for our method and Vit-B.

Transformer in Computer Vision

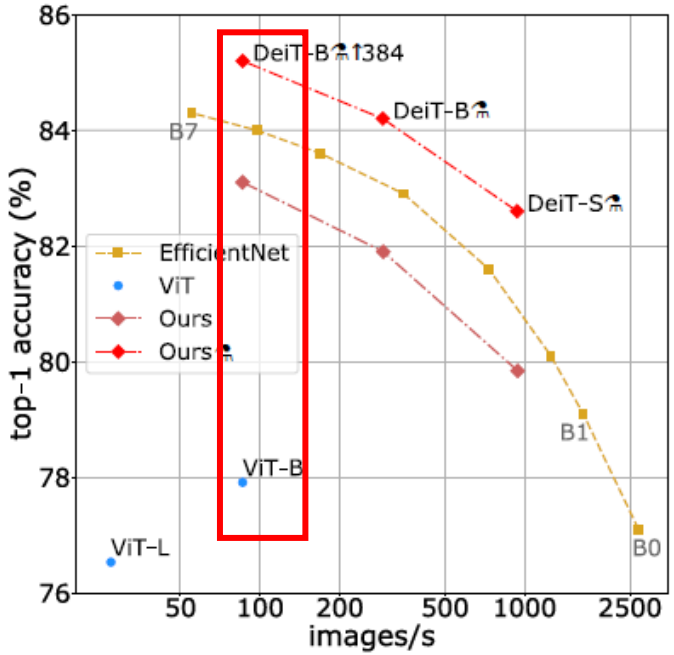
DeiT Experiments

- 효율성 vs 정확도
- ImageNet 데이터셋으로만 학습 진행
- ViT, EfficientNet보다 좋은 성능

Model	#param	Image throughput (image/s)	ImageNet Top-1(ACC)
EfficientNet-B6	66M	96.9	**84.0**
ViT-B/16	86M	85.9	77.9
DeiT-B ↑ 384	86M	85.9	83.1
DeiT-B⚗ ↑ 384	87M	85.8	**84.5**
DeiT-B⚗ ↑ 384/1000epoch	87M	85.8	**85.2**

Data Augmentation

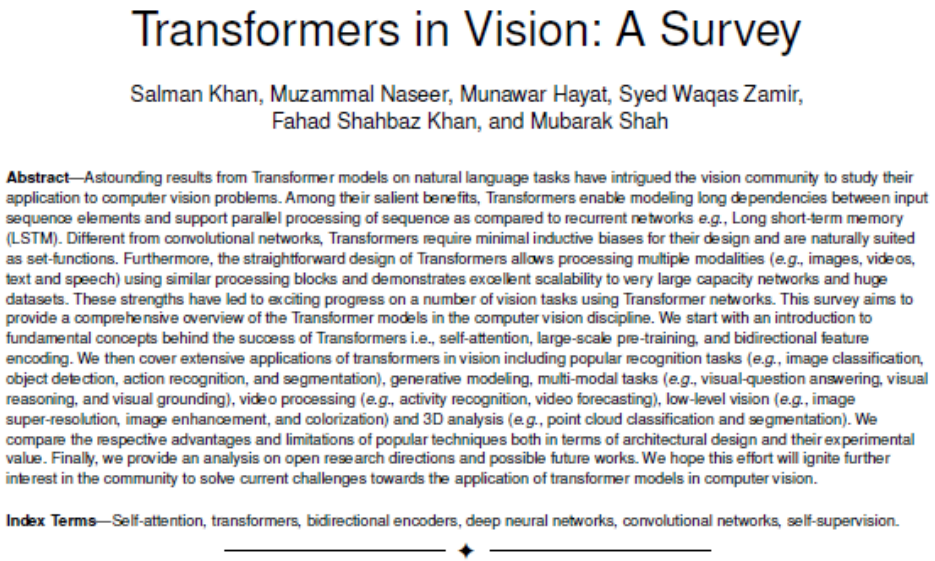
Knowledge Distillation



Conclusion

Transformer in Vision: A Survey

- Computer Vision 분야에서 Transformer가 활용된 연구들 정리한 survey paper
- 관심있는 연구 분야 탐색



Transformers in Vision: A Survey

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir,
Fahad Shahbaz Khan, and Mubarak Shah

Abstract—Astounding results from Transformer models on natural language tasks have intrigued the vision community to study their application to computer vision problems. Among their salient benefits, Transformers enable modeling long dependencies between input sequence elements and support parallel processing of sequence as compared to recurrent networks *e.g.*, Long short-term memory (LSTM). Different from convolutional networks, Transformers require minimal inductive biases for their design and are naturally suited as set-functions. Furthermore, the straightforward design of Transformers allows processing multiple modalities (*e.g.*, images, videos, text and speech) using similar processing blocks and demonstrates excellent scalability to very large capacity networks and huge datasets. These strengths have led to exciting progress on a number of vision tasks using Transformer networks. This survey aims to provide a comprehensive overview of the Transformer models in the computer vision discipline. We start with an introduction to fundamental concepts behind the success of Transformers i.e., self-attention, large-scale pre-training, and bidirectional feature encoding. We then cover extensive applications of transformers in vision including popular recognition tasks (*e.g.*, image classification, object detection, action recognition, and segmentation), generative modeling, multi-modal tasks (*e.g.*, visual-question answering, visual reasoning, and visual grounding), video processing (*e.g.*, activity recognition, video forecasting), low-level vision (*e.g.*, image super-resolution, image enhancement, and colorization) and 3D analysis (*e.g.*, point cloud classification and segmentation). We compare the respective advantages and limitations of popular techniques both in terms of architectural design and their experimental value. Finally, we provide an analysis on open research directions and possible future works. We hope this effort will ignite further interest in the community to solve current challenges towards the application of transformer models in computer vision.

Index Terms—Self-attention, transformers, bidirectional encoders, deep neural networks, convolutional networks, self-supervision.

Task	Method	**Design Highlights** (focus on differences with the standard form)
Image Classification	ViT [11]	Directly adopted NLP Transformer Encoder for images, Mechanism to linearly embed image patches with positional embedding suitable for the Encoder.
	DeiT [12]	Transformer as s student while CNN as a teacher, Distillation tokens to produce estimated labels from teacher, Attention between class and distillation tokens.
	CLIP [81]	Jointly train image and text encoders on image-text pairs, to maximize similarity of valid pairs and minimize otherwise
Object Detection	DETR [13]	Linear projection layer to reduce CNN feature dimension, Spatial positional embedding added to each multi-head self-attention layer of both encoder and decoder. Object queries (output positional encoding) added to each multi-head self-attention layer of decoder.
	D-DETR [14]	Deformable Transformer consists of deformable attention layers to introduce sparse priors in Transformers, Multi-scale attention module.
Low Shot Learning	CT [25]	Self-supervised pretraining, Query-aligned class prototypes that provide spatial correspondence between the support-set images and query image.
Image Colorization	ColTran [24]	Conditional Row/column multi-head attention layers, Progressive multi-scale colorization scheme.
Action Recognition	ST-TR [164]	Spatial and Temporal self-attention to operates on graph data such as joints in skeletons.