CONTENTs

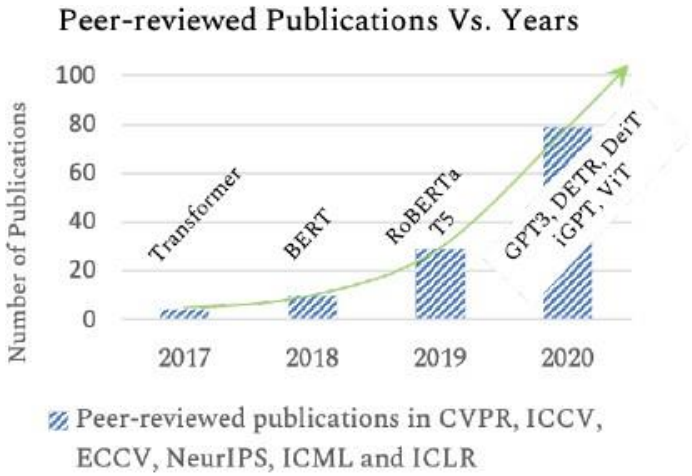
- ◆Introduction
- ◆Transformer
  - Self attention
  - Transformer vs CNN
- ◆Transformer in Computer Vision
  - Vision Transformer (ViT)

※ 출처

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)

⋮

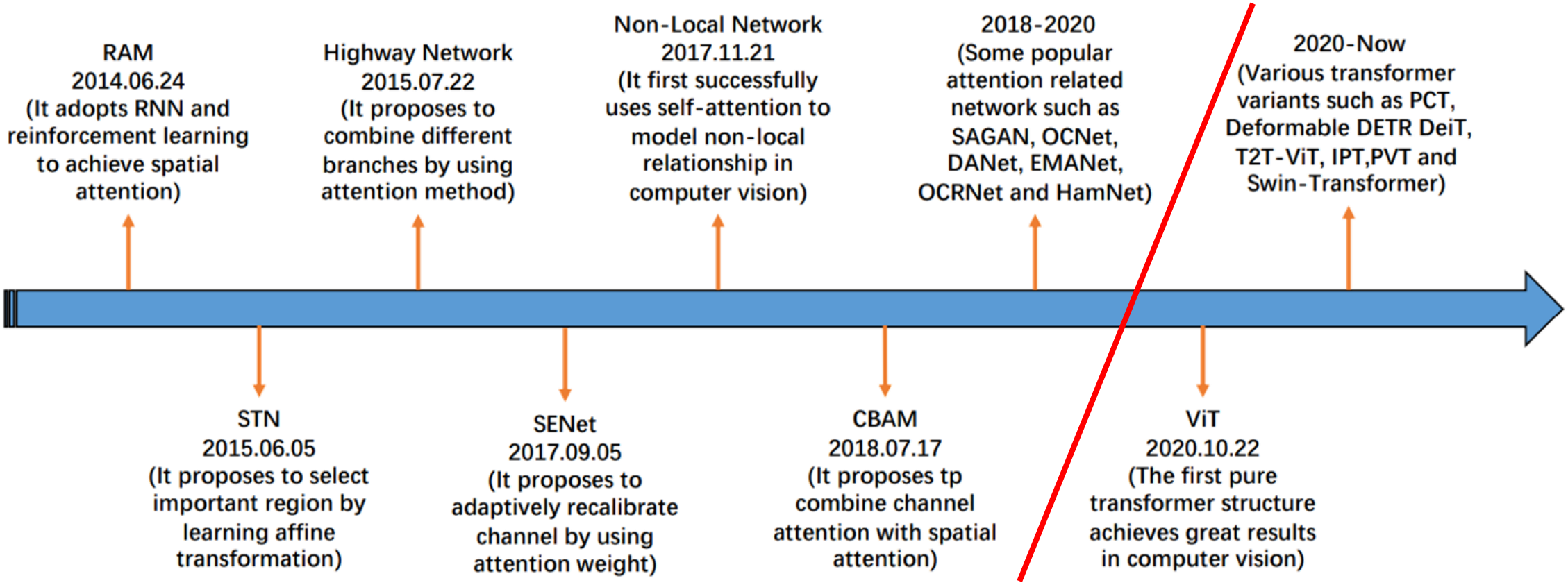
- Vision Transformer (Dosovitskiy et al., 2020)
- Data efficient image Transformer (Touvron et al., 2020)
- TransGAN (Jiang et al., 2021)

⋮

[과거]
CNN**에** self attention**을 어떻게 적용할까?**

[최근]
Transformer 모델 자체를 이용해보자!

Introduction



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: **84,727** (2023.08.11.)

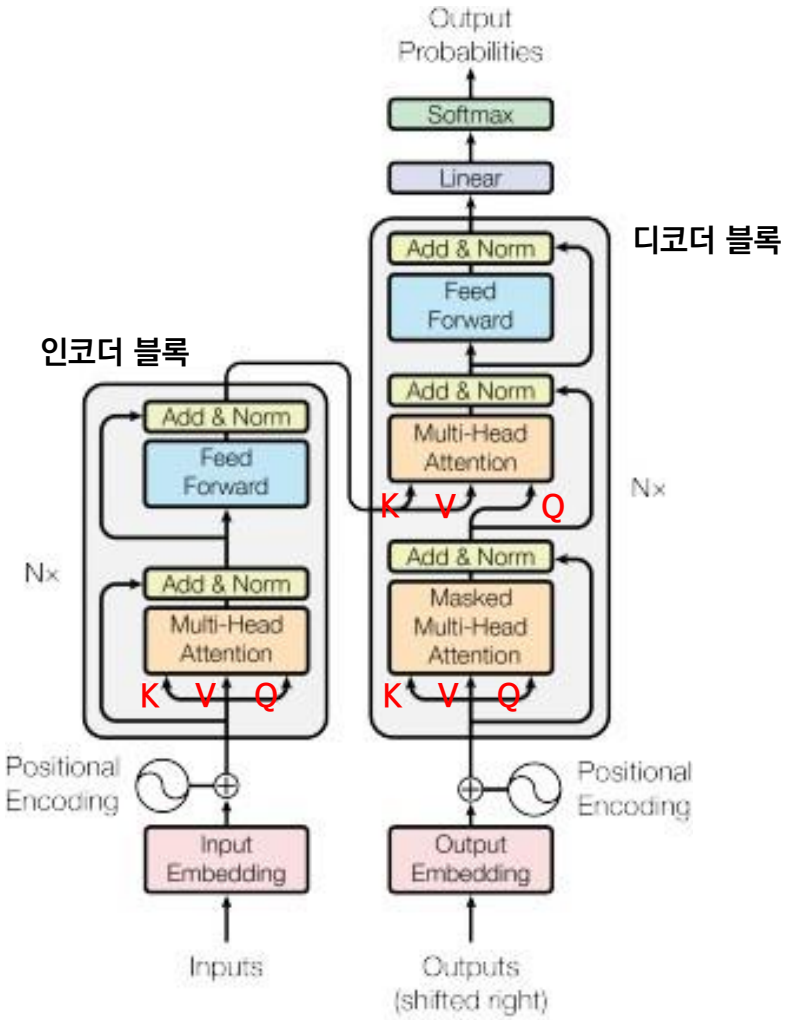
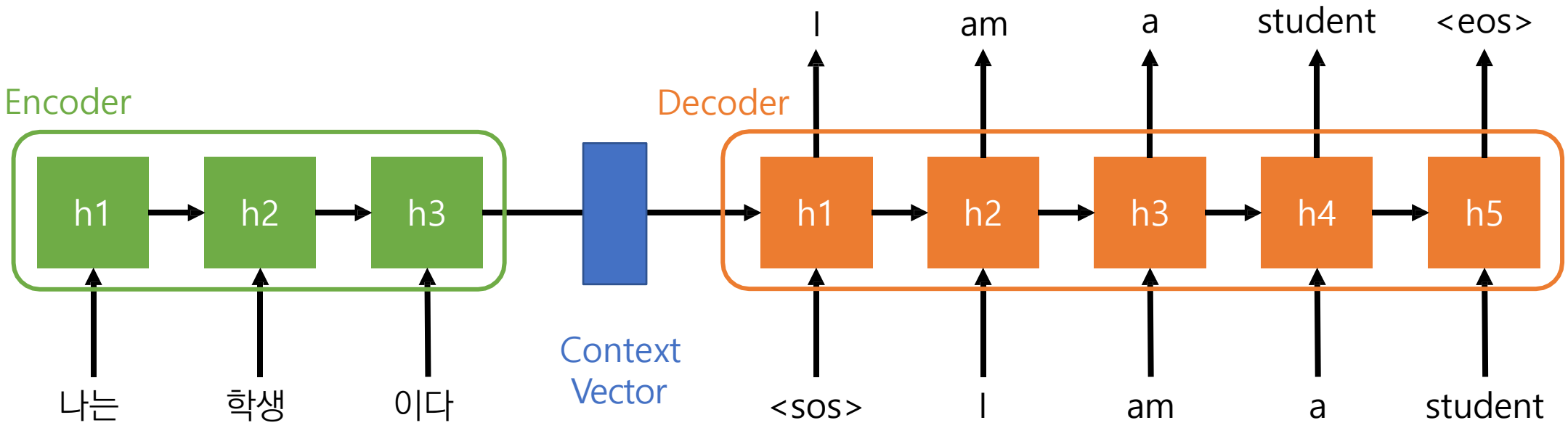


Figure 1: The Transformer - model architecture.

Transformer

Seq2seq

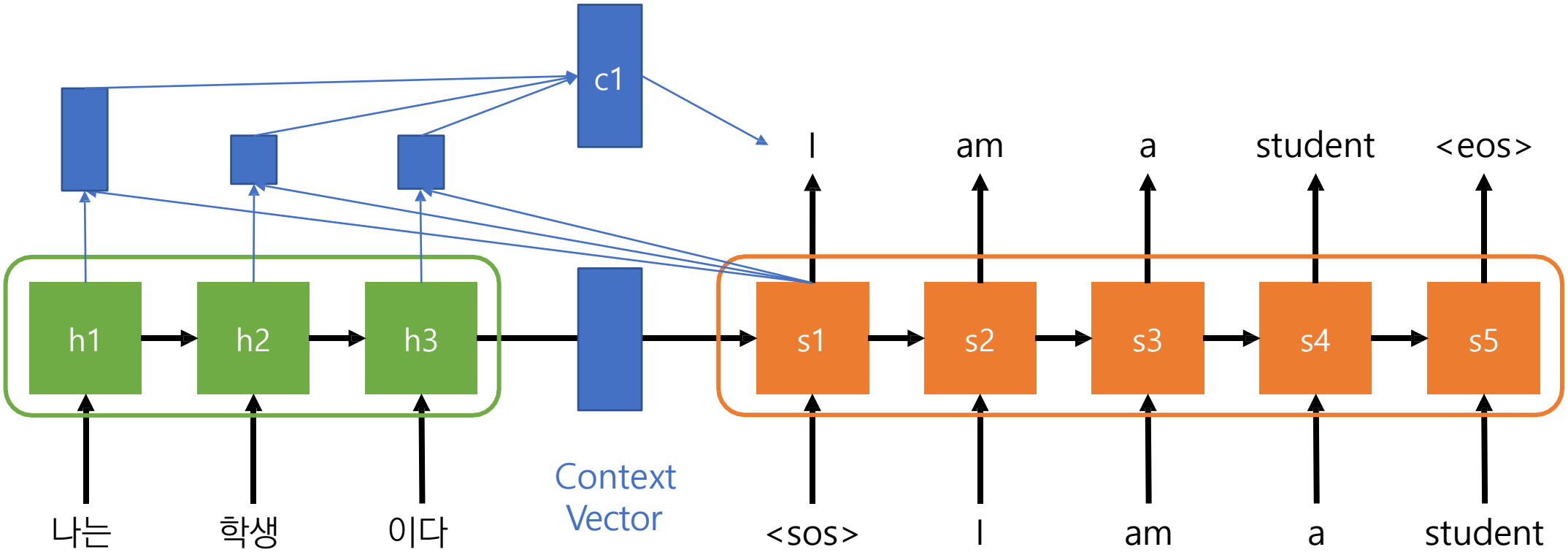
- Seq2seq: 문장을 입력으로 받아 문장을 출력하는 모델 / 기계번역에 주로 사용
- Context vector: Decoder에게 전달되는 입력 문장의 정보
- Context vector의 크기가 제한적이기 때문에 입력 문장의 모든 정보를 전하기 어렵다



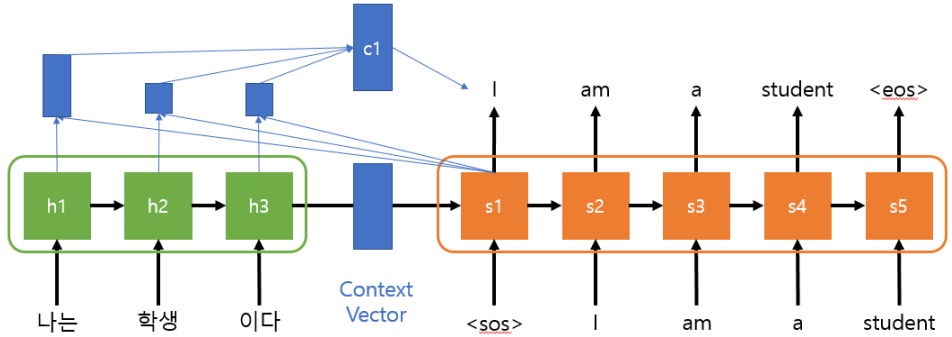
Transformer

Seq2seq with Attention

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

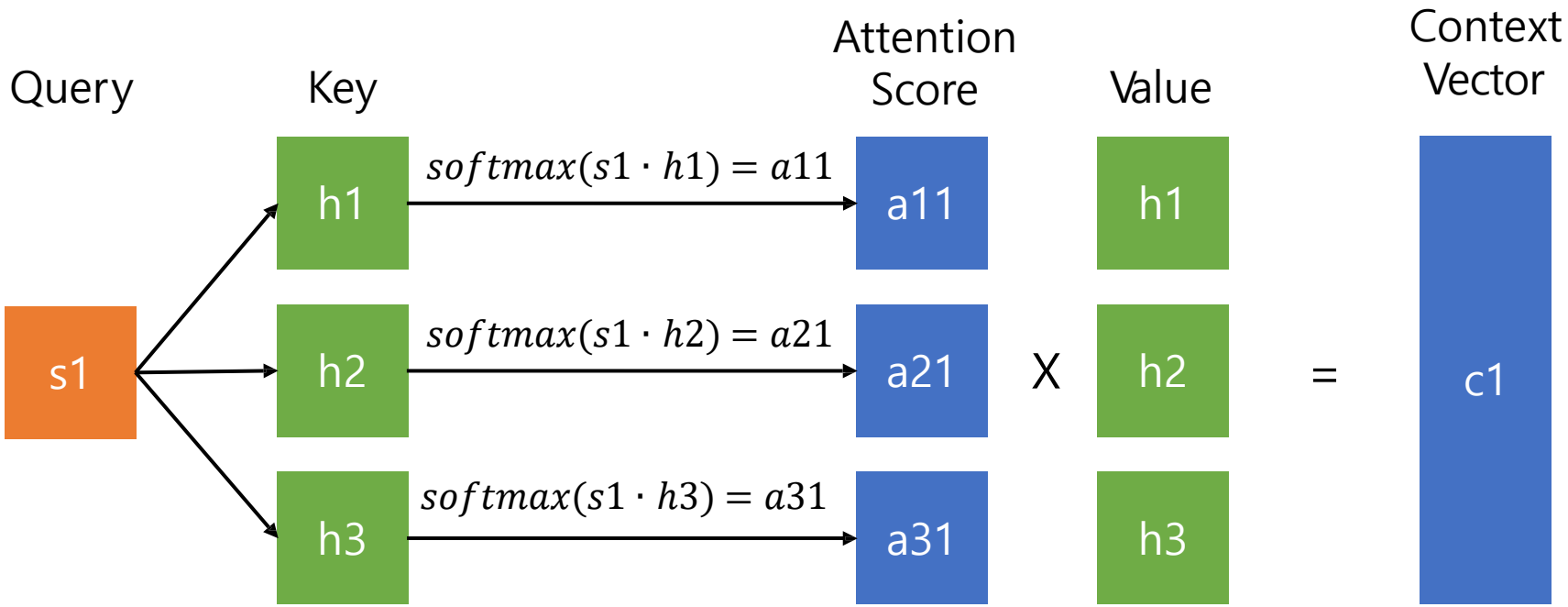


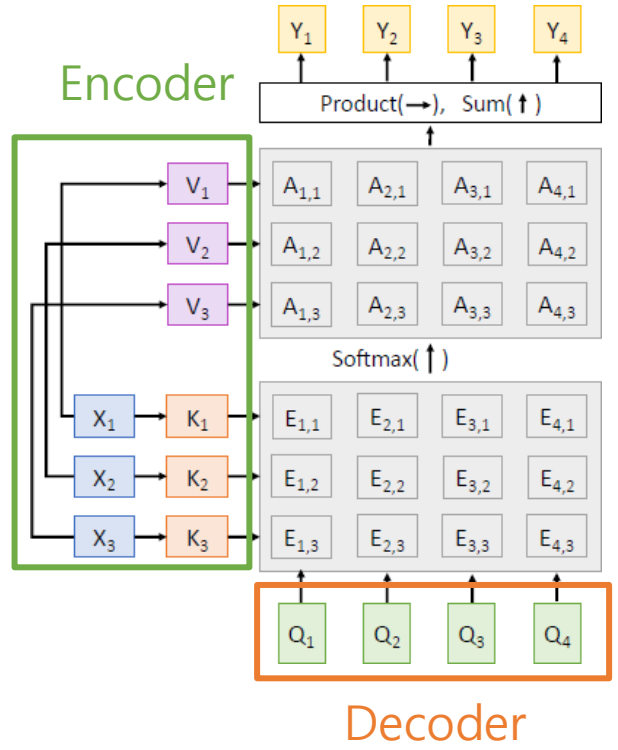
Transformer



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

Seq2seq with Attention

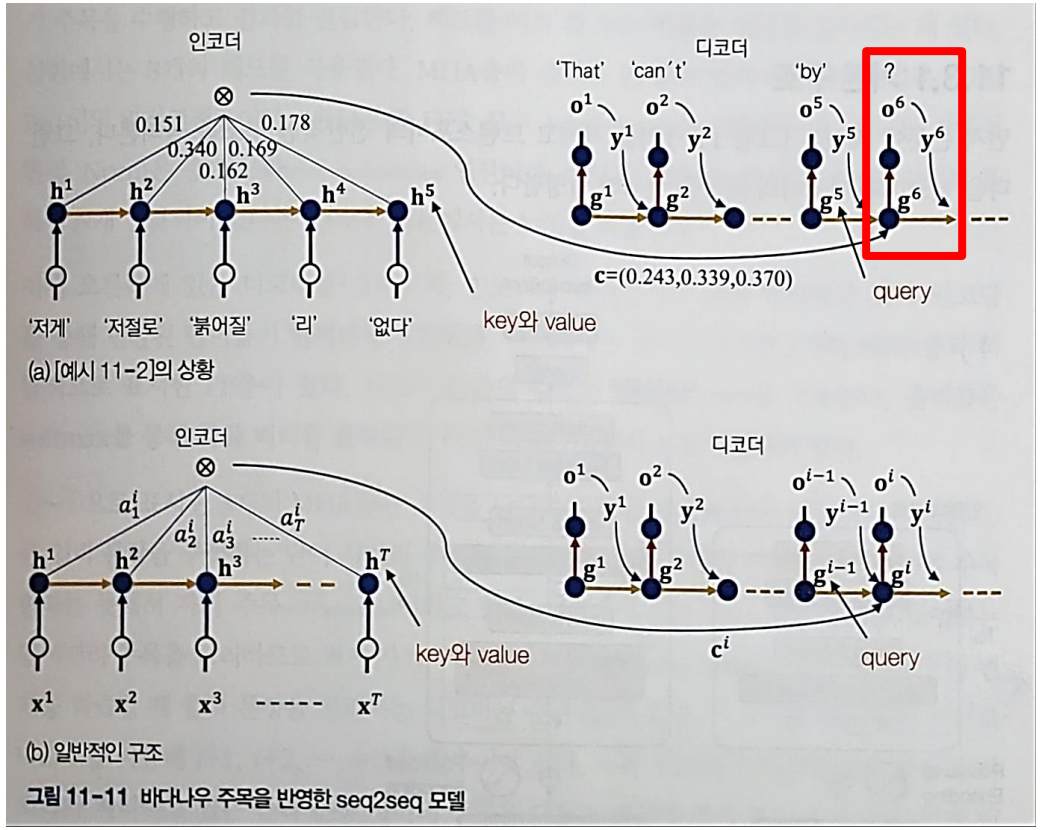


그림 11-11 바다나우 주목을 반영한 seq2seq 모델

$$\mathbf{c} = \text{softmax}(\mathbf{q}\mathbf{K}^{\mathrm{T}})\mathbf{V}$$

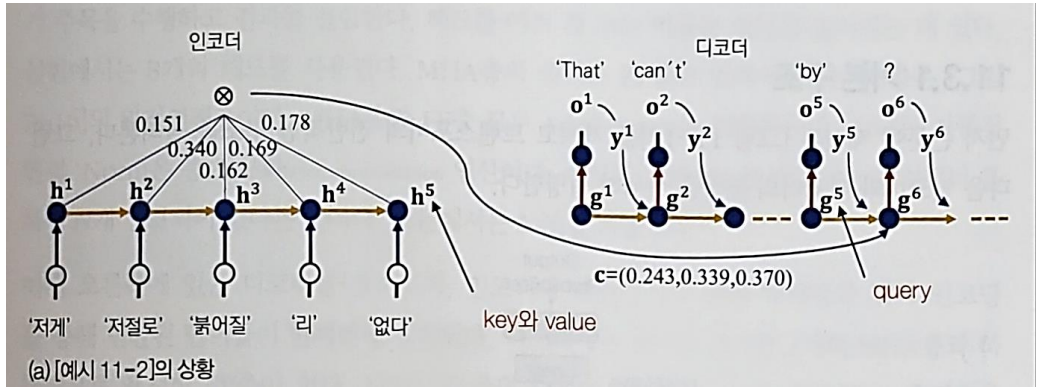
$\mathbf{c}$: Context 벡터
$\mathbf{q}$: Query
$\mathbf{K}$: Key
$\mathbf{V}$: Value
$\text{softmax}(\mathbf{q}\mathbf{K}^{\mathrm{T}})$: Attention 벡터

즉, query가 key와 유사한 정도를 측정하고,
유사도를 가중치로 사용해 value를 가중합 하는 방법!

Transformer

Seq2seq with Attention

"저게 저절로 붉어질 리 없다."



(a) [예시 11-2]의 상황

'저게 저절로 붉어질 리 없다'의 다섯 단어가 [그림 11-11(a)]의 인코더에서 5개의 은닉 상태 $\mathbf{h}^1,\mathbf{h}^2,\cdots,\mathbf{h}^5$를 생성했다. 현재 디코더의 순간 6에 있다고 가정하면 이전 순간 5에서 $\mathbf{g}^5$를 받는다. $\mathbf{g}^5$와 $\mathbf{h}^1 \sim \mathbf{h}^5$가 다음과 같다고 가정한다.

$\mathbf{g}^5=(0.2,0.9,0.0)$

$\mathbf{h}^1=(0.1,0.0,0.8)$ $\mathbf{h}^2=(0.1,0.9,0.0)$ $\mathbf{h}^3=(0.0,0.1,0.8)$ $\mathbf{h}^4=(0.2,0.1,0.6)$ $\mathbf{h}^5=(0.9,0.0,0.1)$

이때 $\mathbf{g}^5$가 query이고, $\mathbf{h}^1,\mathbf{h}^2,\cdots,\mathbf{h}^5$가 key와 value다. 이러한 결정을 행렬로 표현하면 다음과 같다.

$$\mathbf{q}=\begin{pmatrix}0.2 & 0.9 & 0.0\end{pmatrix},\ \mathbf{K}=\begin{pmatrix}0.1 & 0.0 & 0.8\\ 0.1 & 0.9 & 0.0\\ 0.0 & 0.1 & 0.8\\ 0.2 & 0.1 & 0.6\\ 0.9 & 0.0 & 0.1\end{pmatrix},\ \mathbf{V}=\begin{pmatrix}0.1 & 0.0 & 0.8\\ 0.1 & 0.9 & 0.0\\ 0.0 & 0.1 & 0.8\\ 0.2 & 0.1 & 0.6\\ 0.9 & 0.0 & 0.1\end{pmatrix}$$

$$\mathbf{c}=\text{softmax}\left(\mathbf{q}\mathbf{K}^T\right)\mathbf{V}=\text{softmax}\left(\begin{pmatrix}0.2 & 0.9 & 0.0\end{pmatrix}\begin{pmatrix}0.1 & 0.1 & 0.0 & 0.2 & 0.9\\ 0.0 & 0.9 & 0.1 & 0.1 & 0.0\\ 0.8 & 0.0 & 0.8 & 0.6 & 0.1\end{pmatrix}\right)\begin{pmatrix}0.1 & 0.0 & 0.8\\ 0.1 & 0.9 & 0.0\\ 0.0 & 0.1 & 0.8\\ 0.2 & 0.1 & 0.6\\ 0.9 & 0.0 & 0.1\end{pmatrix}$$

$$=\begin{pmatrix}0.151 & 0.340 & 0.162 & 0.169 & 0.178\end{pmatrix}\begin{pmatrix}0.1 & 0.0 & 0.8\\ 0.1 & 0.9 & 0.0\\ 0.0 & 0.1 & 0.8\\ 0.2 & 0.1 & 0.6\\ 0.9 & 0.0 & 0.1\end{pmatrix}$$

$$=\begin{pmatrix}0.243 & 0.339 & 0.370\end{pmatrix}$$

Transformer

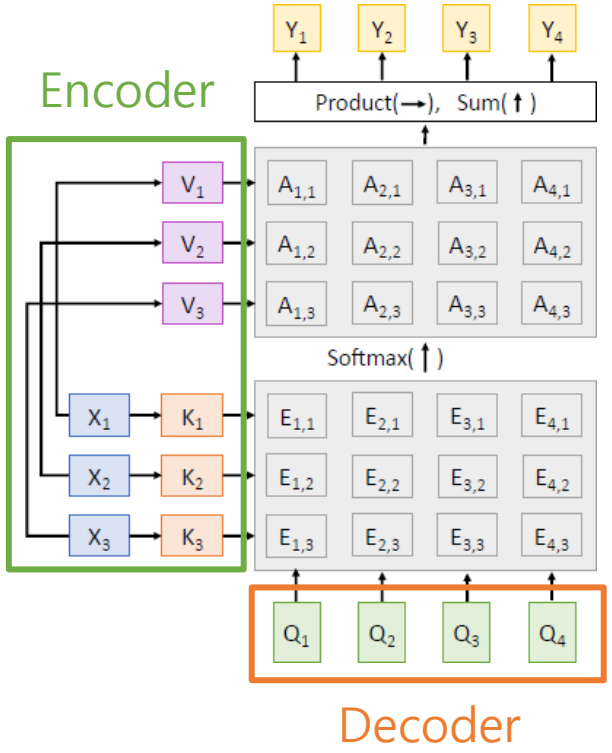
Attention vs Self Attention

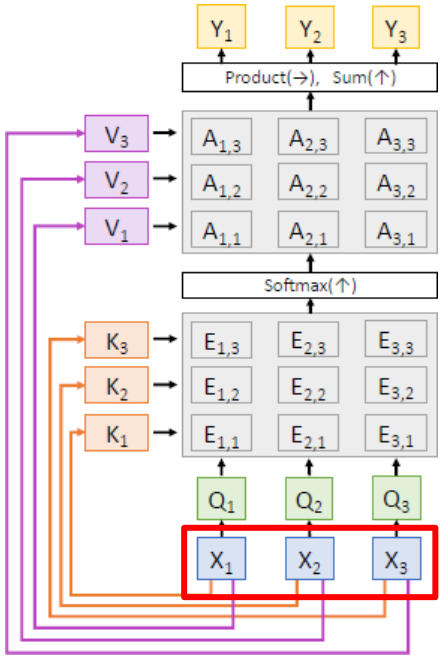
- Attention : (Decoder → Query / Encoder → Key, Value) / encoder, decoder 사이의 상관관계를 바탕으로 특징 추출
- Self-Attention : (입력 데이터 → Query, Key, Value) / 데이터 내의 상관관계를 바탕으로 특징 추출

"저게 저절로 붉어질 리 없다."

↓

"That can't turn red by itself."





"영희가 신호등을 보고 길을 건너는데,
그게 금방 노란색으로 변했다."

Input Data

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

Transformer

Attention vs Self Attention

Attention

$$\mathbf{c} = \text{softmax}(\mathbf{q}\mathbf{K}^{\mathrm{T}})\mathbf{V}$$

$\mathbf{c}$: Context 벡터

$\mathbf{q}$: Query

$\mathbf{K}$: Key

$\mathbf{V}$: Value

$\text{softmax}(\mathbf{q}\mathbf{K}^{\mathrm{T}})$: Attention 벡터

query (디코더)가 key (인코더)와 유사한 정도를 측정하고,
유사도를 가중치로 사용해 value (인코더)를 가중합 하는 방법!

Self-Attention
(트랜스포머)

$$\mathbf{C} = \text{softmax}(\mathbf{Q}\mathbf{K}^{\mathrm{T}})\mathbf{V}$$

$\mathbf{C}$: Context 행렬

$\mathbf{Q} = \mathbf{X}\mathbf{W}^{\mathrm{Q}}$: Query

$\mathbf{K} = \mathbf{X}\mathbf{W}^{\mathrm{K}}$: Key

$\mathbf{V} = \mathbf{X}\mathbf{W}^{\mathrm{V}}$: Value

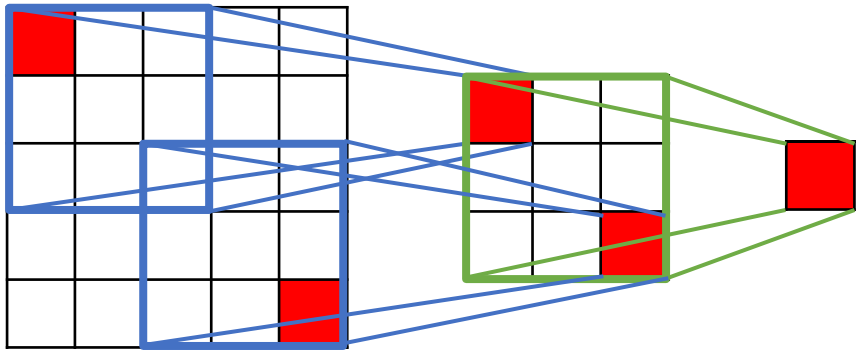
$\text{softmax}(\mathbf{Q}\mathbf{K}^{\mathrm{T}})$: Attention 행렬

**즉, query (X)가 key (X)와 유사한 정도를 측정하고,
유사도를 가중치로 사용해 value (X)를 가중합 하는 방법!**

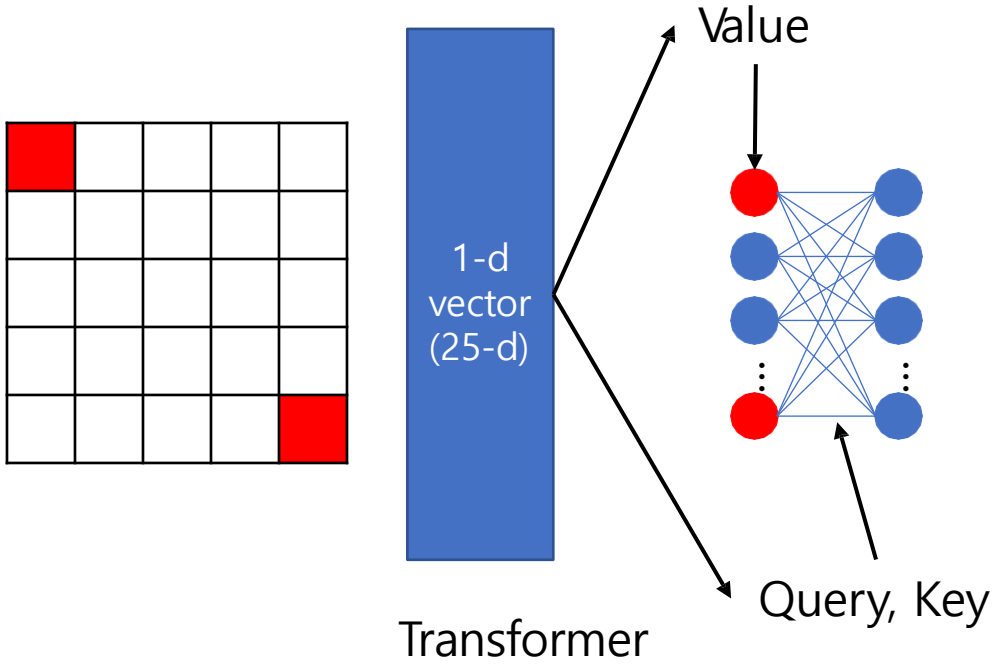
Transformer

Transformer vs CNN

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



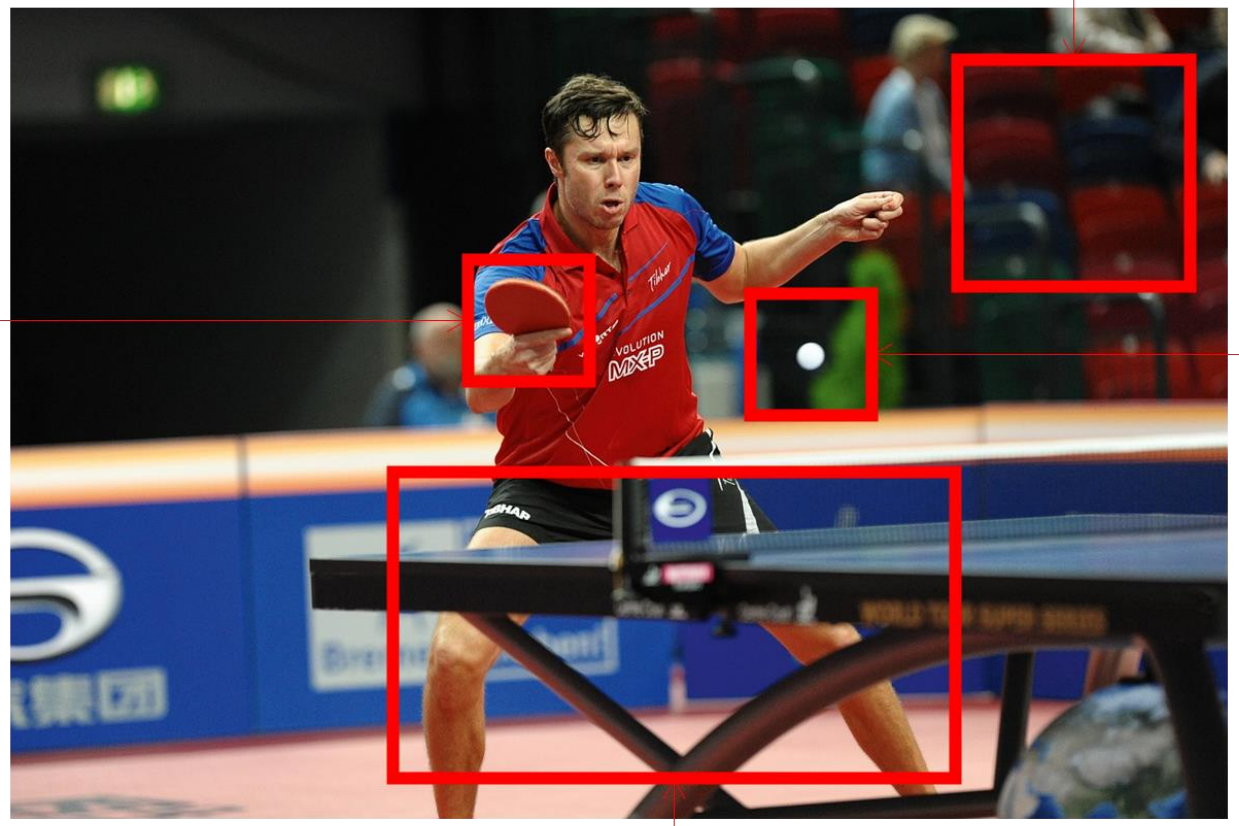
CNN



Transformer

Transformer in Computer Vision

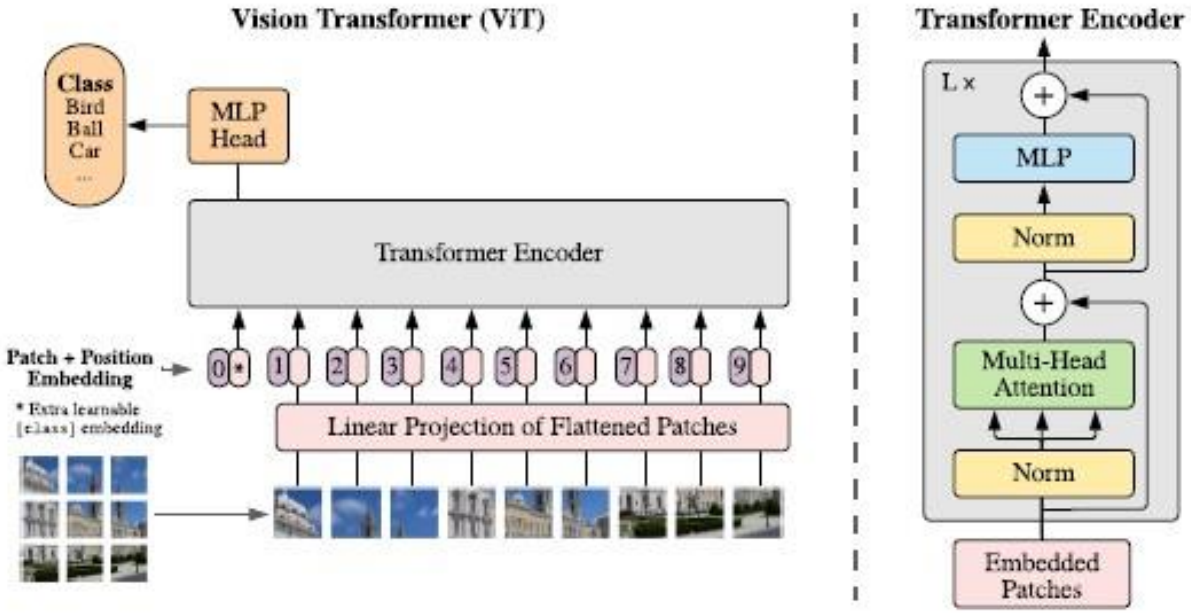
Vision Transformer (ViT)



Transformer in Computer Vision

Vision Transformer (ViT)

- **19,457**회 인용 (23.08.11. 기준)
- Transformer 구조를 활용해 image classification을 수행
- 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

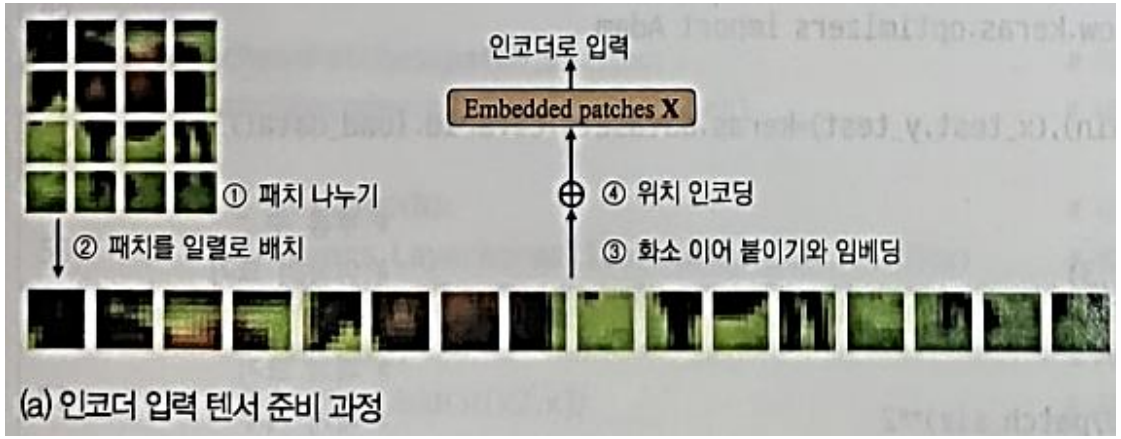
{adosovitskiy, neilhoulsby}@google.com

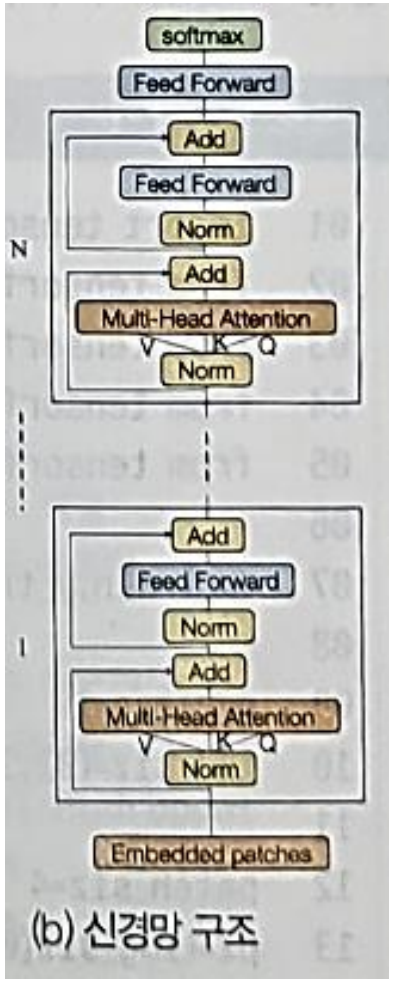
ABSTRACT

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]

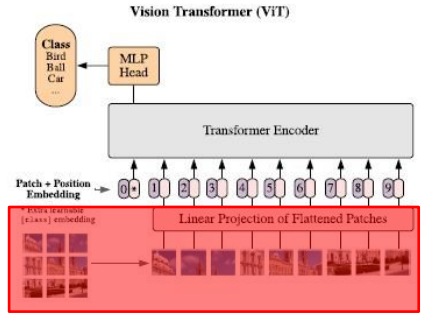
Transformer in Computer Vision

Vision Transformer (ViT)





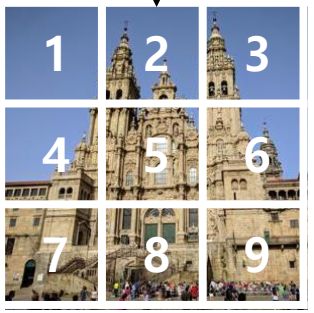
Transformer in Computer Vision



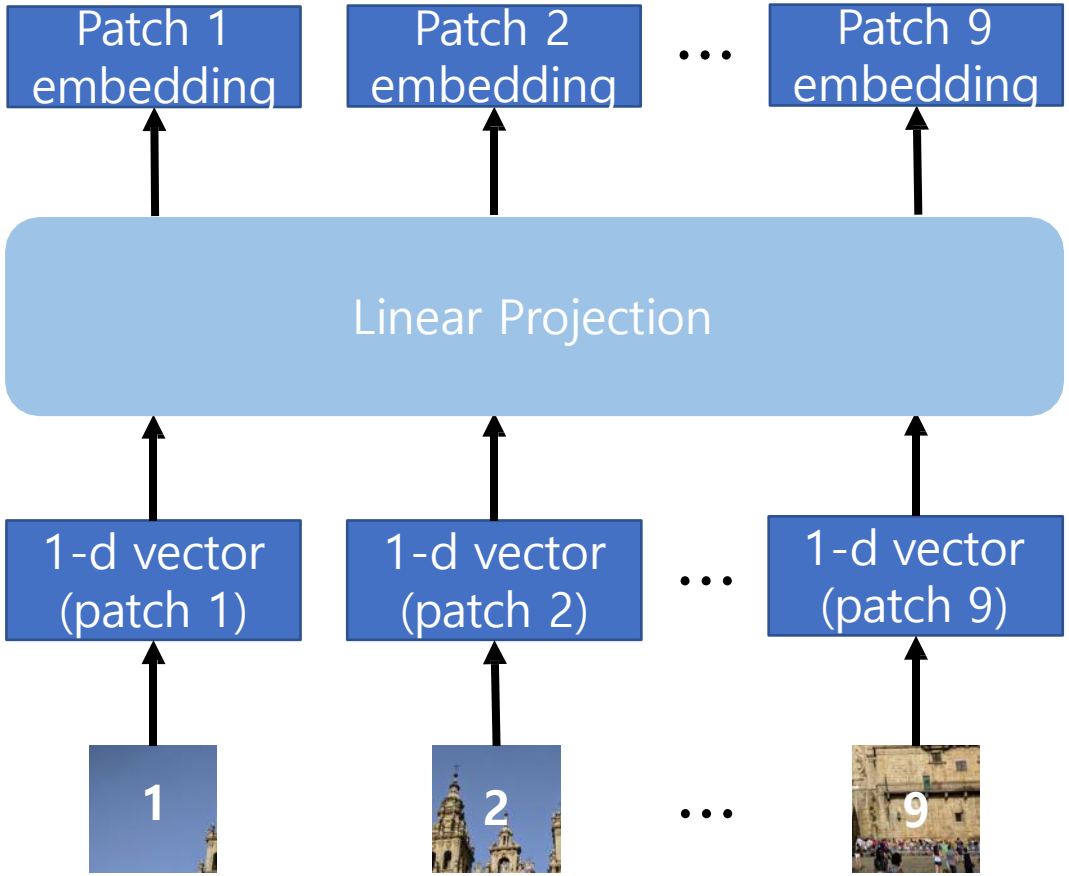
Vision Transformer example

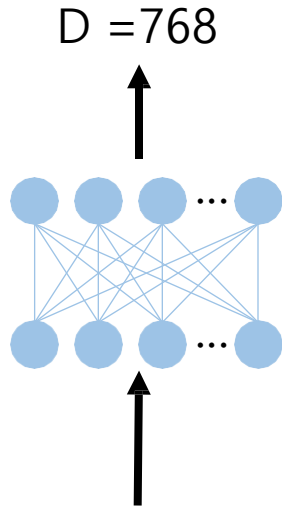
Input Image

48X48

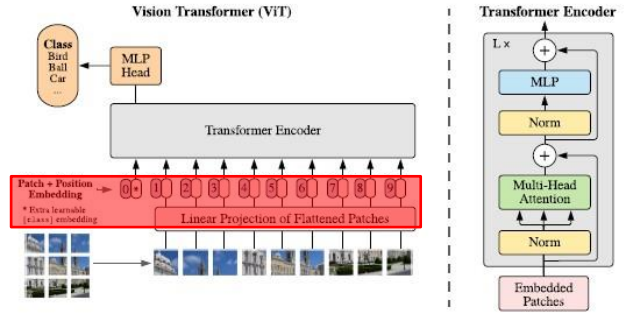


(16X16)*9



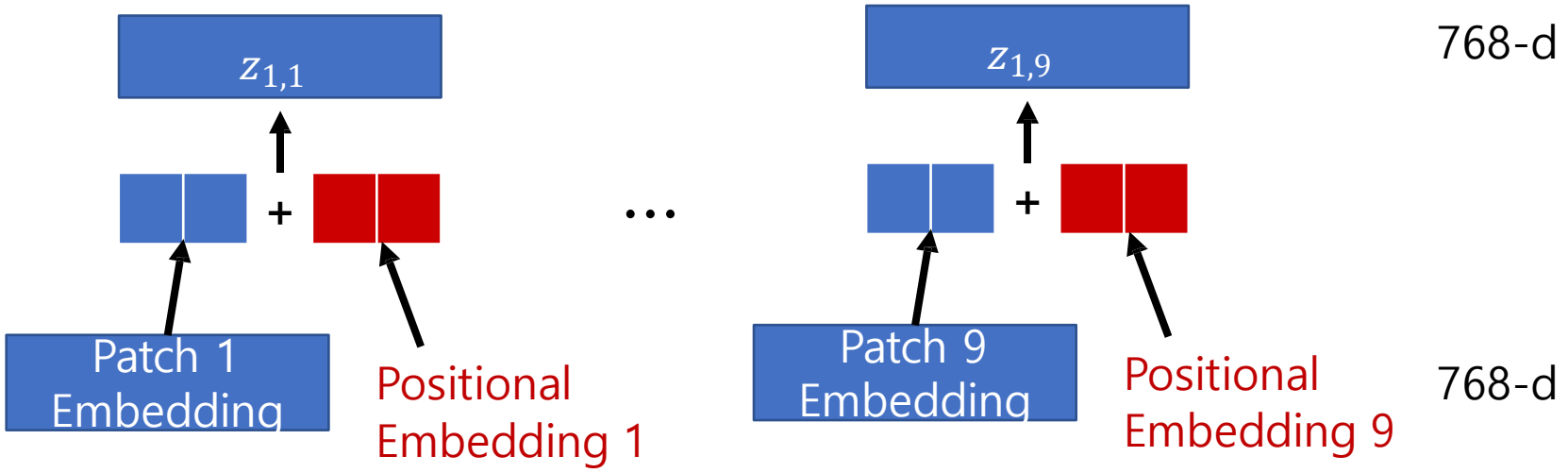


Transformer in Computer Vision

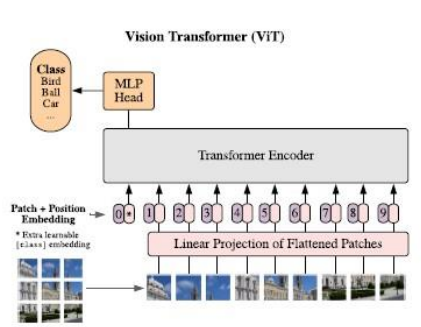


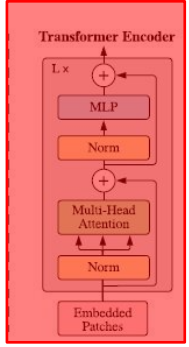
Vision Transformer example

- Position embedding: patch의 위치 정보 (학습을 통해 결정)
- Patch embedding + Positional embedding = Transformer encoder 입력



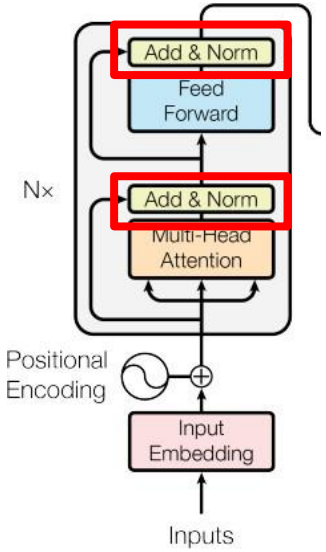
Transformer in Computer Vision



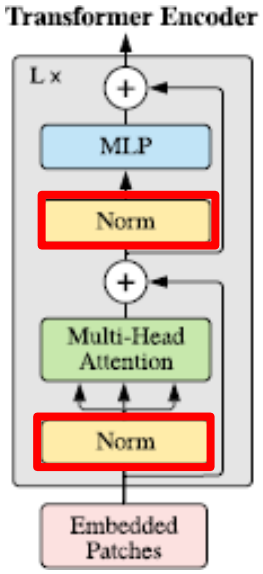


Vision Transformer example

- "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

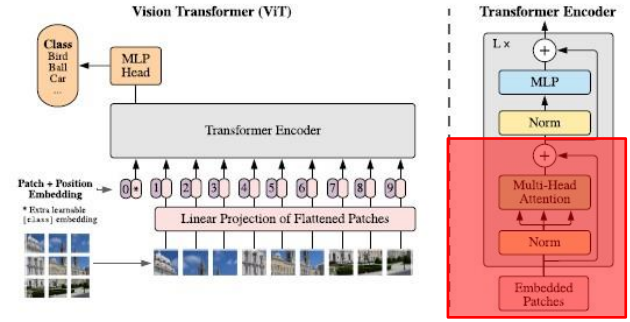


"**ViT**" Transformer

Learning Deep Transformer Models for Machine Translation

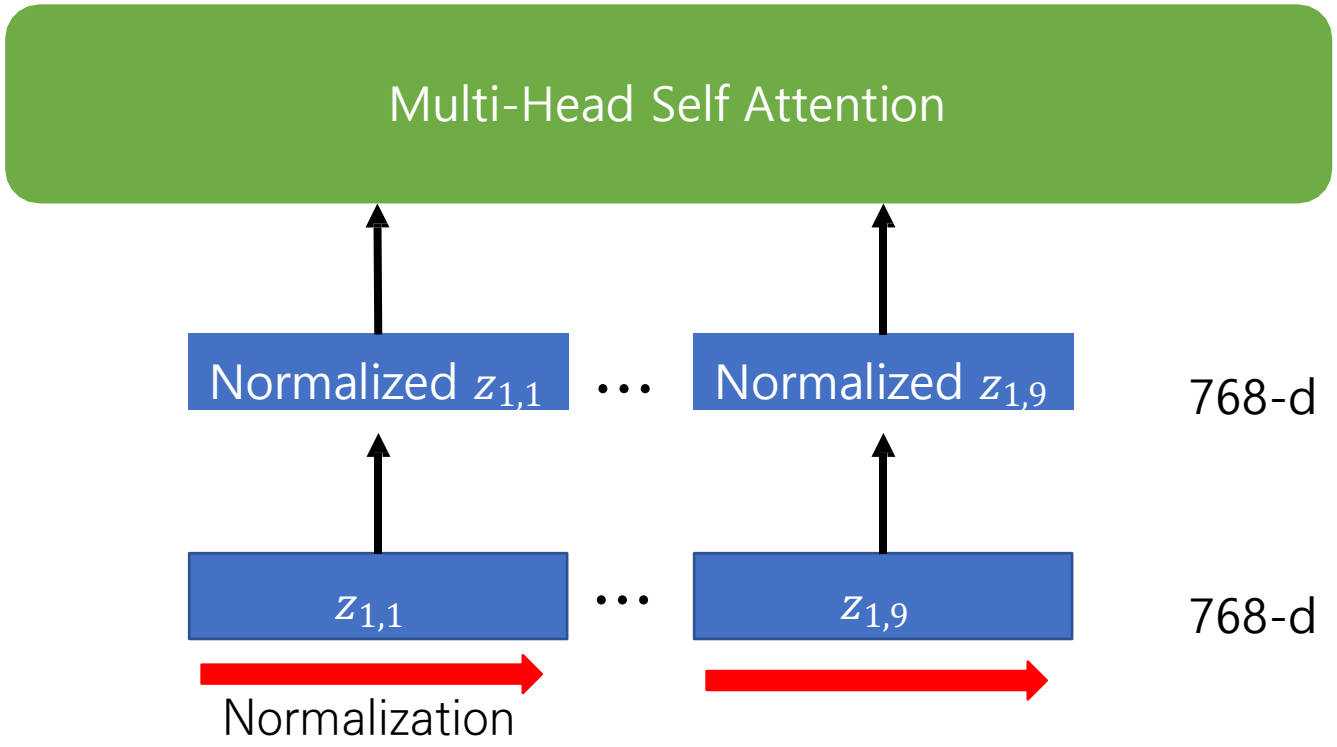
Qiang Wang[1], Bei Li[1], Tong Xiao[1,2]*, Jingbo Zhu[1,2], Changliang Li[3], Derek F. Wong[4], Lidia S. Chao[4]
[1]NLP Lab, Northeastern University, Shenyang, China
[2]NiuTrans Co., Ltd., Shenyang, China
[3]Kingsoft AI Lab, Beijing, China
[4]NLP²CT Lab, University of Macau, Macau, China
wangqiangneu@gmail.com,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

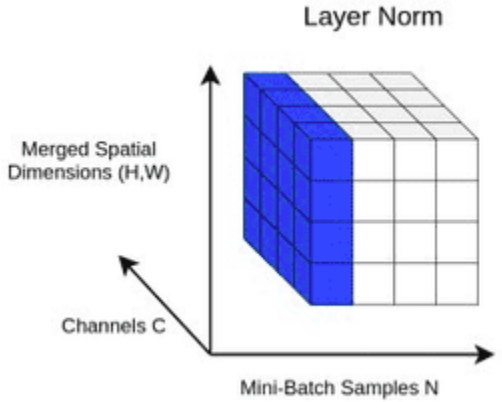
Transformer in Computer Vision

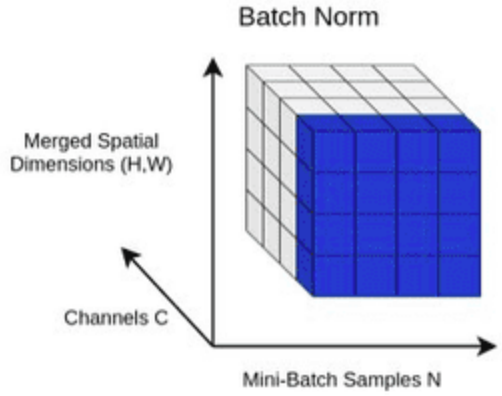


Vision Transformer example

- Transformer encoder: Layer normalization

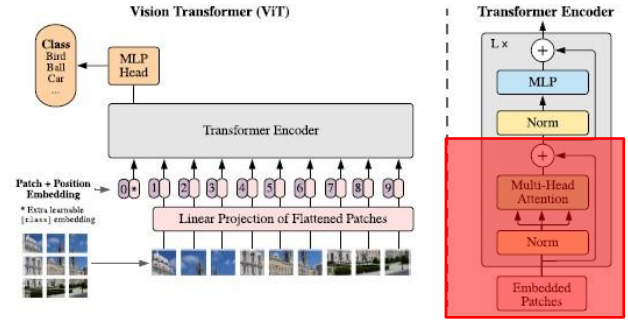






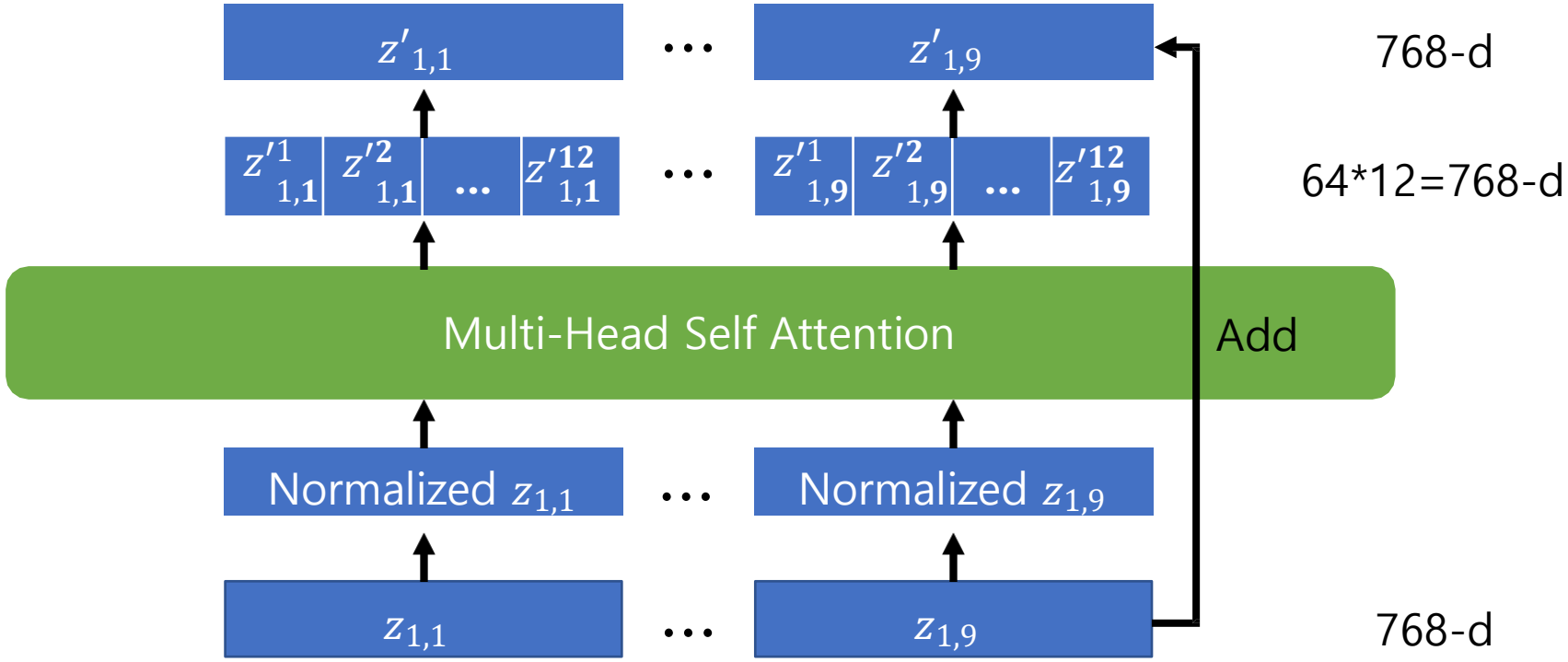
https://theaisummer.com/normalization/

Transformer in Computer Vision

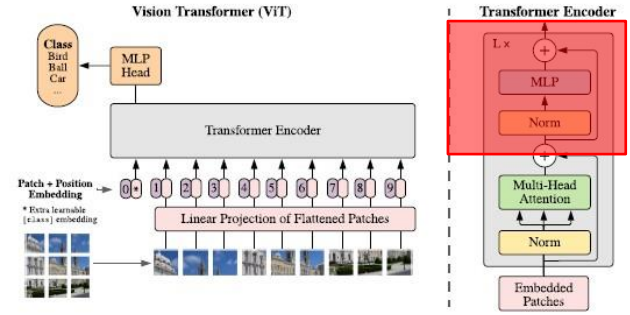


Vision Transformer example

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

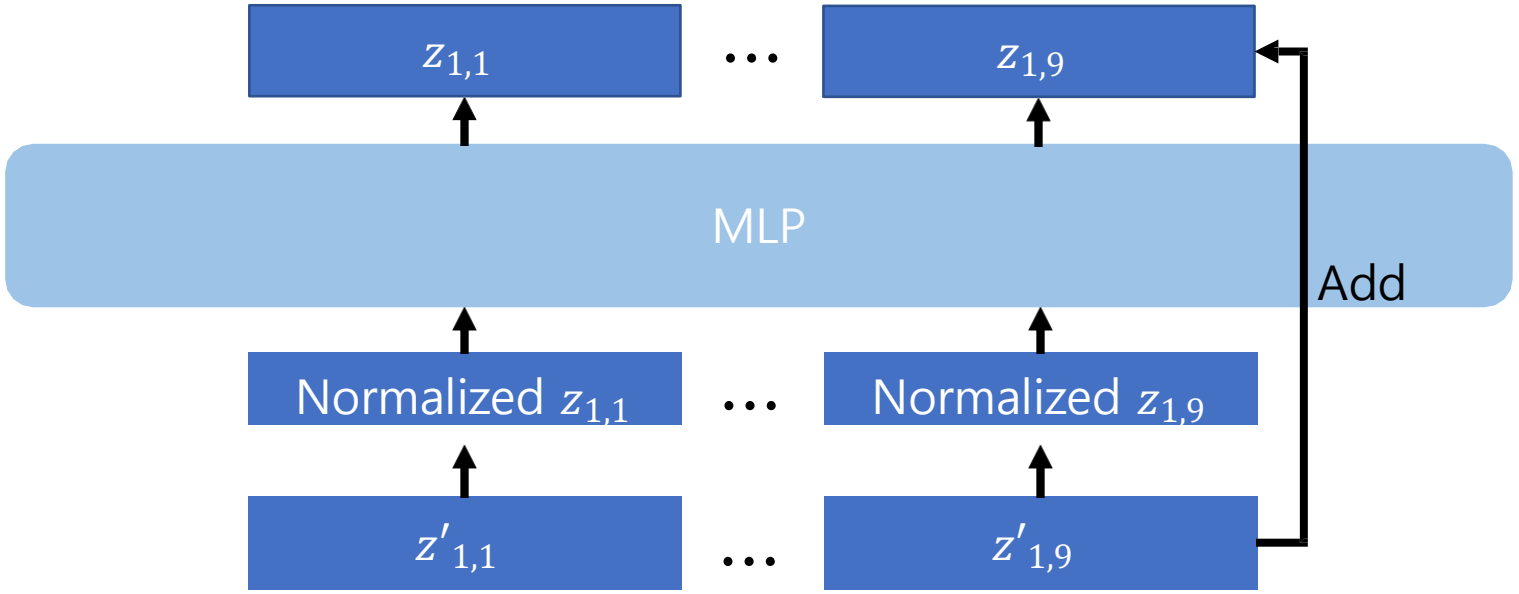


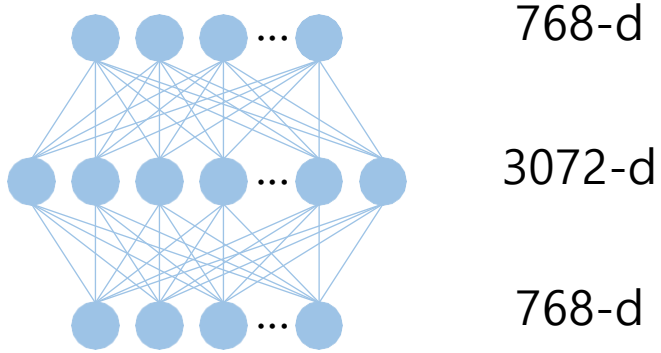
Transformer in Computer Vision



Vision Transformer example

- Transformer encoder: MLP





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

※ TPUv3-core-days : 하루 기준 사용된 TPU core 개수

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

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