

---

# How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers

---

2022.01.21

Data Mining and Quality Analytics Lab

박진혁

# 목차

---

1. Introduction
2. What is ViT?
3. How to train your ViT?
4. Conclusion

# 1. Introduction

---

## ❖ 발표자 소개



- 박진혁
- Data Mining & Quality Analytics Lab(김성범 교수님)
- 석.박사 통합과정 5학기 재학 중(2019.8 ~)

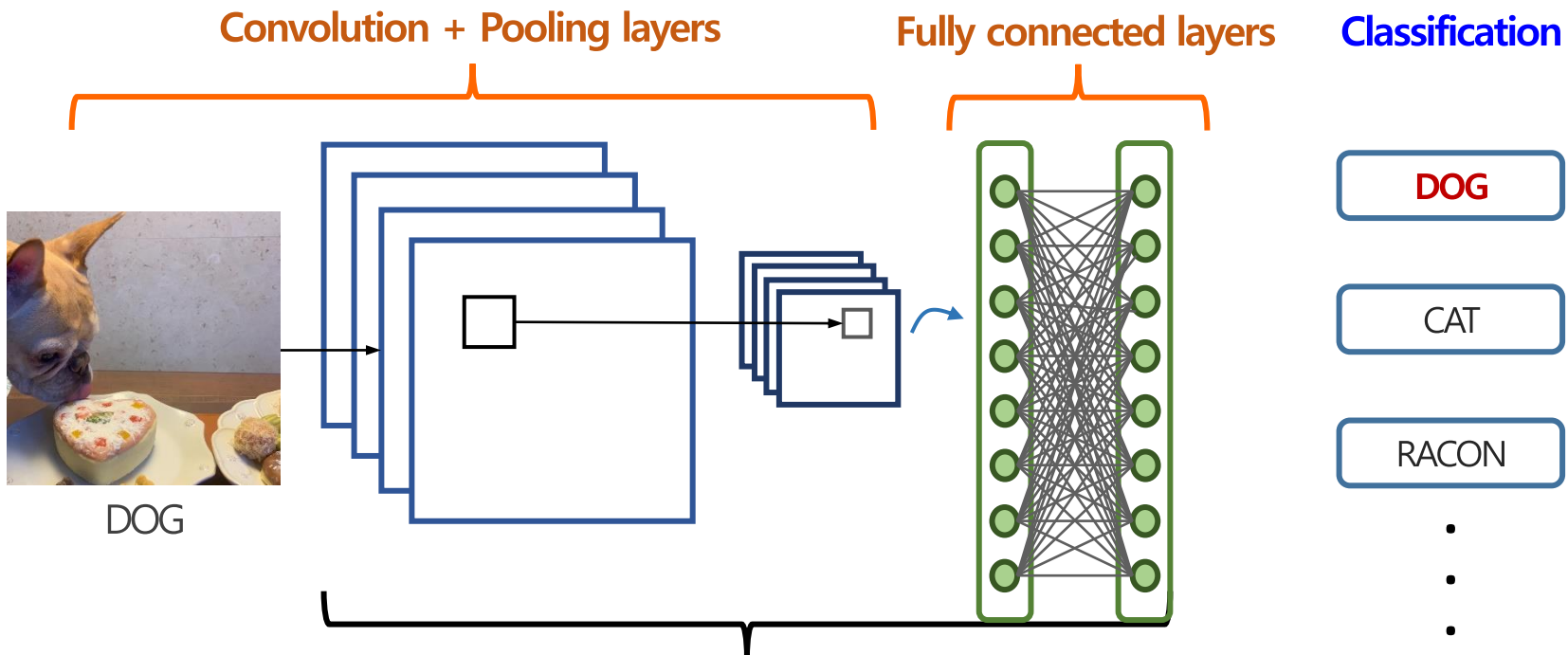
## ✓ 관심 분야

- Computer Vision
- Image Captioning
- Multimodal Learning

# 1. Introduction

## ❖ From CNN to ViT

- Convolutional Neural Network
  - Computer vision분야에서 가장 많이 사용되는 architecture
  - 이미지를 입력 받아 이미지의 공간정보를 유지한 채 학습함

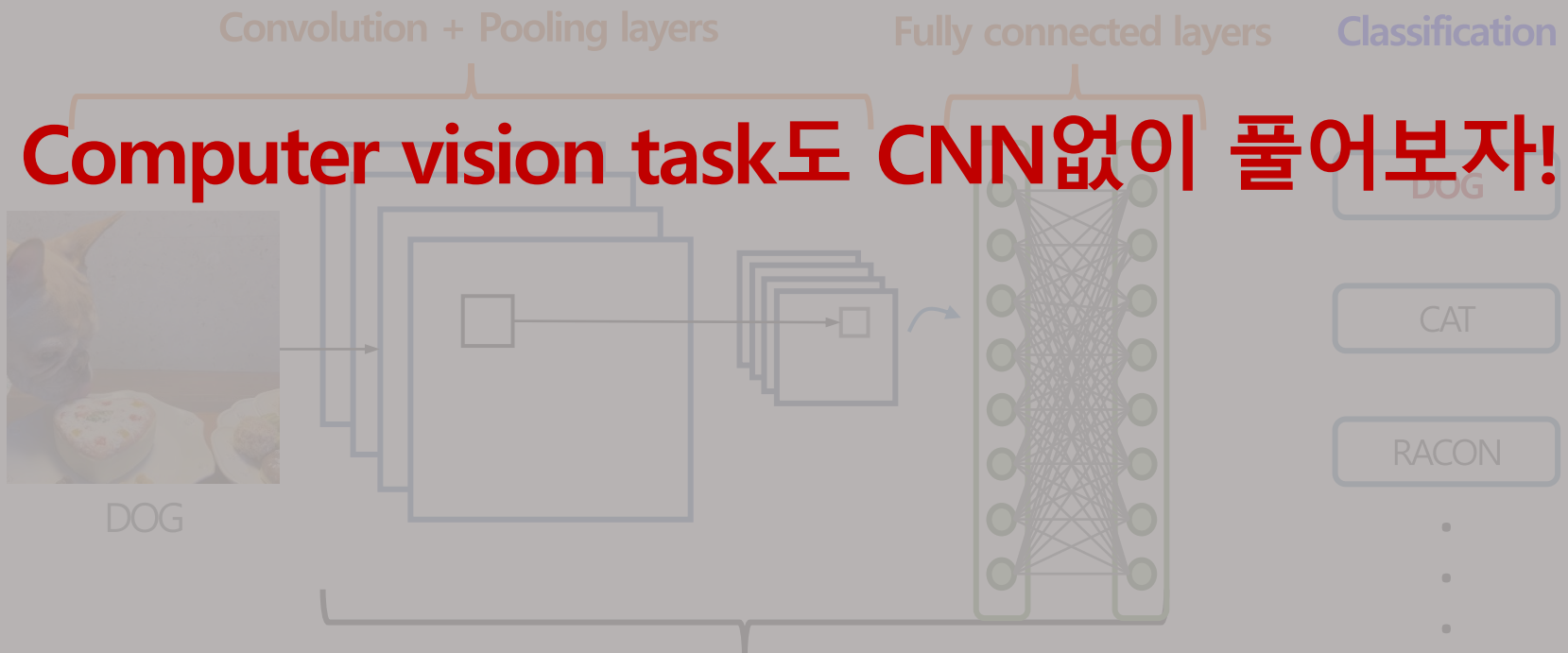


# 1. Introduction

## ❖ From CNN to ViT

### ➤ Convolutional Neural Network

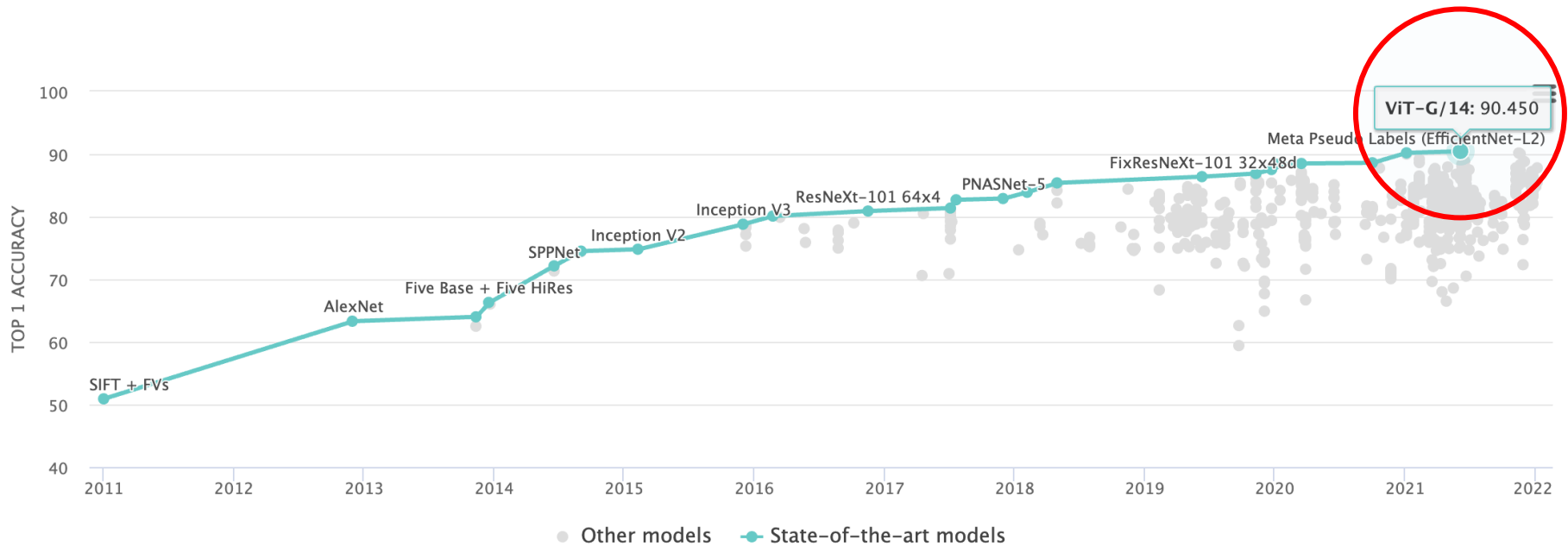
- Computer vision분야에서 가장 많이 사용되는 architecture
- 이미지를 입력 받아 이미지의 공간정보를 유지한 채 학습함



# 1. Introduction

## ❖ From CNN to ViT

- Computer vision분야에서 합성곱 신경망이 아닌 Transformer를 적용
- Transformer architecture가 computer vision분야에서 수 많은 SOTA를 달성
- Computer vision분야에 NLP architecture의 적용 가능성을 보여준 논문



## 2. What is ViT?

---

### ❖ Vision Transformer(ViT)

- 2022년 1월 17일 기준 2196회 인용
- Google Research에서 발표
- Transformer architecture를 활용하여 image classification을 수행

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

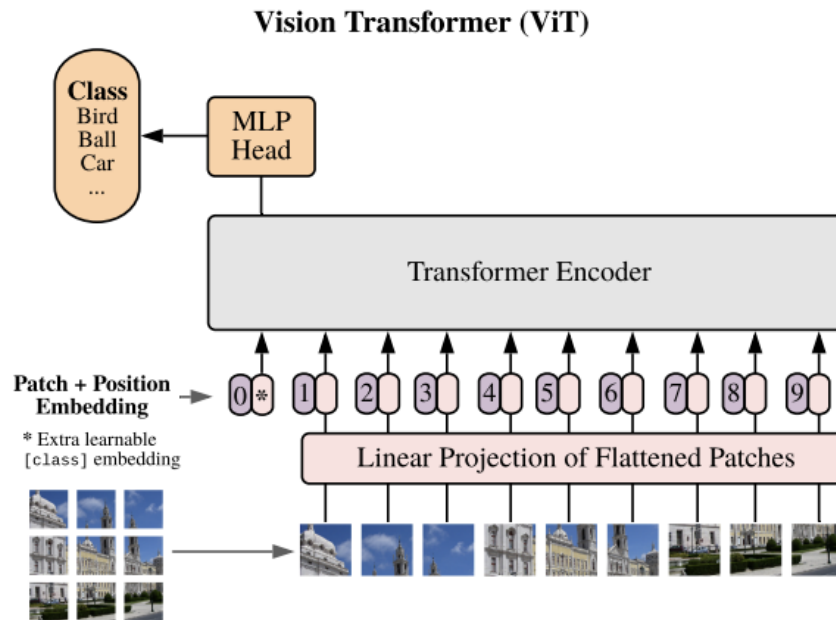
**Alexey Dosovitskiy<sup>\*,†</sup>, Lucas Beyer<sup>\*</sup>, Alexander Kolesnikov<sup>\*</sup>, Dirk Weissenborn<sup>\*</sup>,  
Xiaohua Zhai<sup>\*</sup>, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer,  
Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby<sup>\*,†</sup>**

<sup>\*</sup>equal technical contribution, <sup>†</sup>equal advising  
Google Research, Brain Team

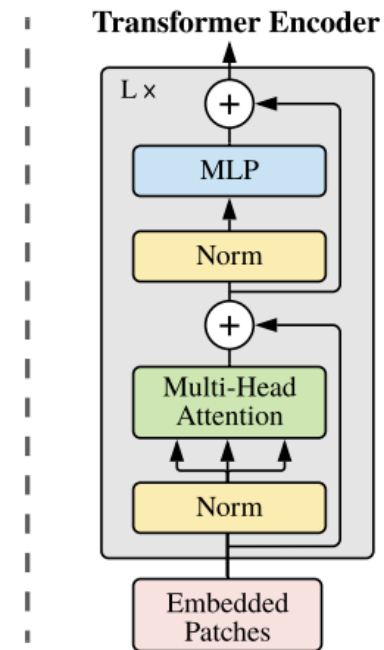
## 2. What is ViT?

### ❖ Vision Transformer(ViT)

- 입력 이미지
- 모델 아키텍처
- 대용량 데이터의 사전학습



<Model overview>

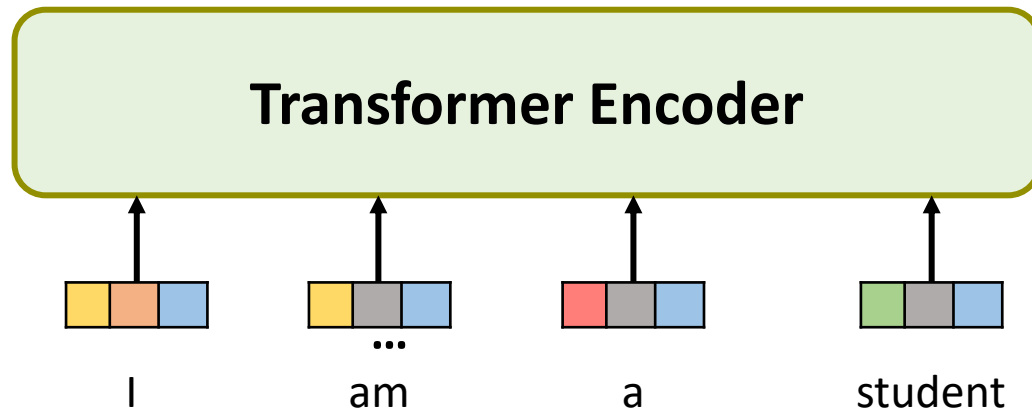




## 2. What is ViT?

### ❖ Vision Transformer(ViT)

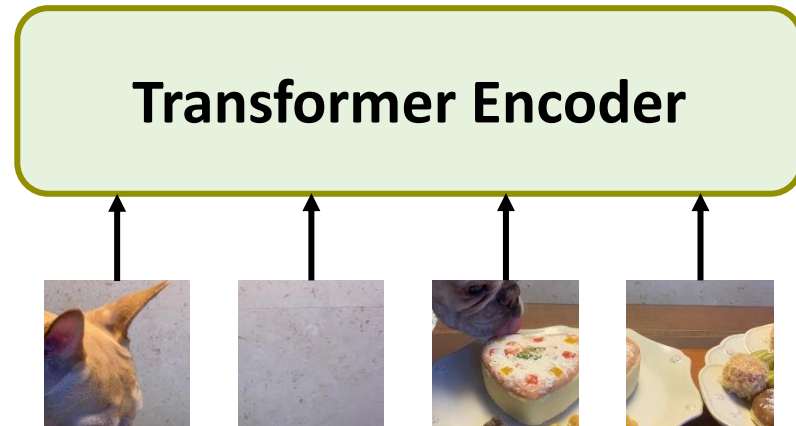
- 입력 이미지
  - 기존 Transformer에서는 token embedding을 입력
  - 이미지를 patch형태로 나눠서 입력
  - 모든 patch를 벡터와 flatten하게 생성한 뒤 입력



## 2. What is ViT?

### ❖ Vision Transformer(ViT)

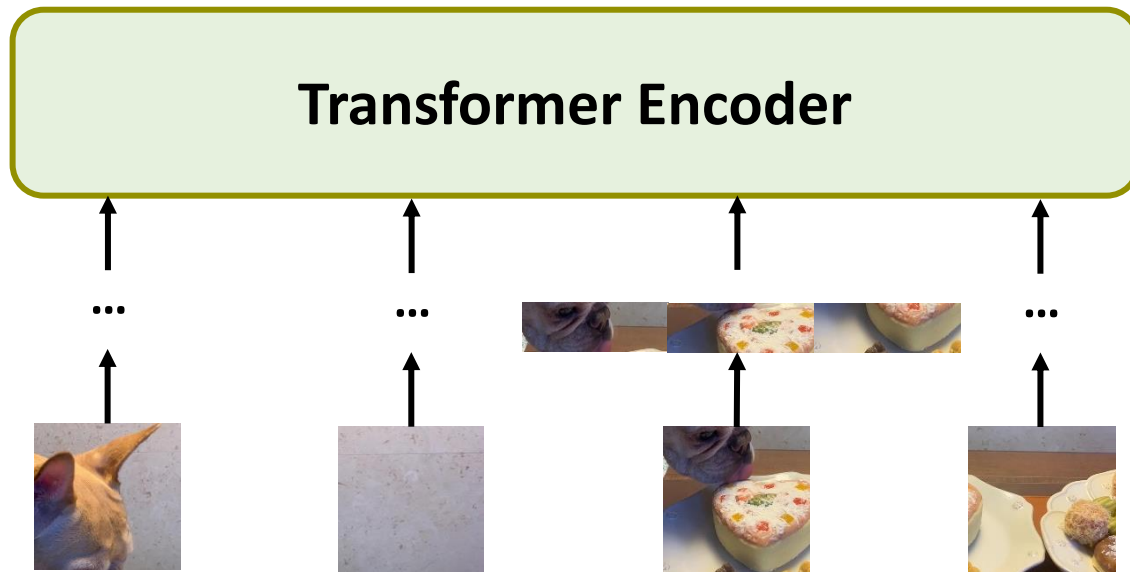
- 입력 이미지
  - 기존 Transformer에서는 token embedding을 입력
  - 이미지를 patch형태로 나눠서 입력
  - 모든 patch를 벡터와 flatten하게 생성한 뒤 입력



## 2. What is ViT?

### ❖ Vision Transformer(ViT)

- 입력 이미지
  - 기존 Transformer에서는 token embedding을 입력
  - 이미지를 patch형태로 나눠서 입력
  - 모든 patch를 벡터와 flatten하게 생성한 뒤 입력

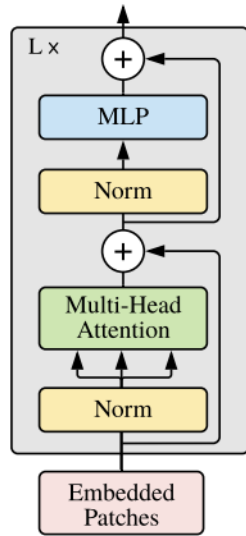


## 2. What is ViT?

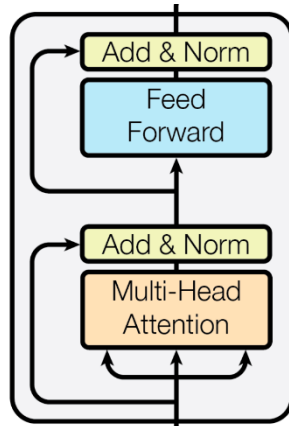
### ❖ Vision Transformer(ViT)

- 모델 아키텍처
  - 기존 Transformer의 encoder부분과 거의 유사함
  - Transformer와 차이점: Norm의 위치, GELU사용
- 대용량 데이터의 사전학습
  - 대용량 데이터셋으로 사전학습을 한 뒤 downstream task로 fine-tuning을 진행

Transformer Encoder



<ViT>



<Transformer>

私は学生です

。



Je suis étudiant.

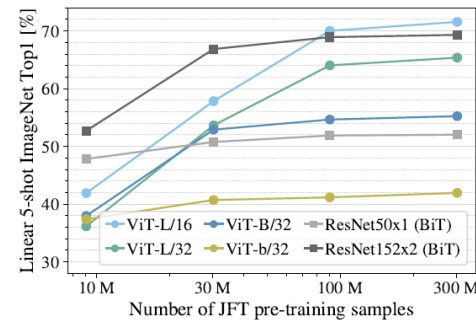
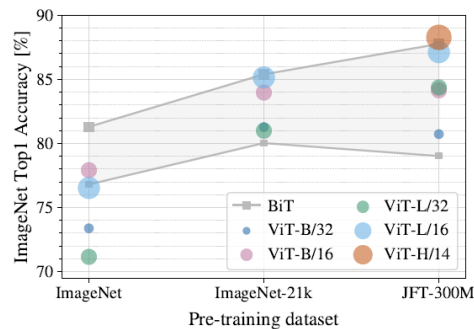
## 2. What is ViT?

### ❖ Vision Transformer(ViT)

#### ➤ Result

- JFT로 사전학습을 진행한 뒤, 각 데이터 셋에 Transfer Learning진행
- CNN기반의 SoTA모델과 비교해서 ViT가 SoTA를 갱신함
- 사전학습을 진행할 경우 데이터 셋이 클수록 ViT성능이 향상
  - Convolutional Inductive Bias가 존재하지 않기 때문에 많은 데이터가 필요

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

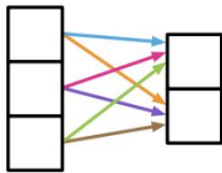


## 2. What is ViT?

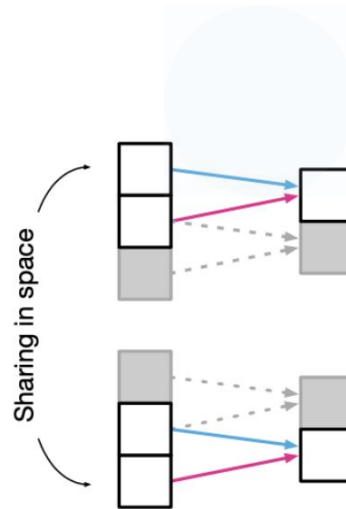
### ❖ Vision Transformer(ViT)

#### ➤ Inductive bias

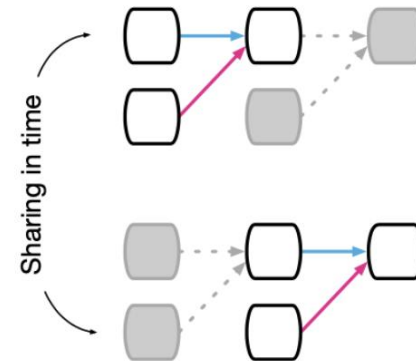
- 학습 모델이 지금까지 만나보지 못했던 상황에서 정확한 예측을 하기 위해 사용하는 추가적인 가정
- Fully Connected
- Convolutional
- Recurrent



(a) Fully connected



(b) Convolutional



(c) Recurrent

## 2. What is ViT?

<http://dmqa.korea.ac.kr/activity/seminar/316>

## 논문

# Neural Information Processing Systems (Neural IPS)에서 발표된 논문

the Brain과 Google Research 그룹에서 발표한 논문

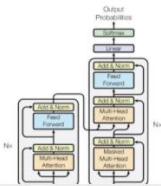
- 2020년 9월 3일 기준으로 약 11600회 인용

### Attention Is All You Need


<b>Ashish Vaswani*</b> Google Brain <a href="mailto:avaswani@google.com">avaswani@google.com</a>	<b>Noam Shazeer*</b> Google Brain <a href="mailto:nshazeer@google.com">nshazeer@google.com</a>	<b>Niki Parmar*</b> Google Research <a href="mailto:niki14@google.com">niki14@google.com</a>	<b>Jacob Devlin*</b> Google Research <a href="mailto:devlinj@google.com">devlinj@google.com</a>
<b>Łukasz Kaiser*</b> Google Research <a href="mailto:lkaiser@google.com">lkaiser@google.com</a>	<b>Aleksandr N. Senior†</b> University of Toronto <a href="mailto:senior@cs.toronto.edu">senior@cs.toronto.edu</a>	<b>Colin Raffel*</b> Google Brain <a href="mailto:colinr@google.com">colinr@google.com</a>	

**Will Fedus\***  
[willf@google.com](mailto:willf@google.com)

**Abstract**  
 The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that inhibit an explicit notion of parallelism. We present a new simple network architecture, [called Transformer](#),



## Transformer



발표자: **이영재**

📅

2020년 9월 4일

🕒

오후 1시 ~

▶

온라인 비디오 시청 (YouTube)

## 세미나 정보 보기 →

<http://dmqa.korea.ac.kr/activity/seminar/295>

### 3. How to train your ViT?

---

- ❖ How To Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers
  - 2022년 1월 17일 기준 24회 인용
  - Google Research에서 발표
  - ViT model을 효율적으로 학습시키는 방법론에 대한 비교분석 실험 진행

---

## How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers

---

**Andreas Steiner\*, Alexander Kolesnikov\*, Xiaohua Zhai\***

**Ross Wightman<sup>†</sup>, Jakob Uszkoreit, Lucas Beyer\***

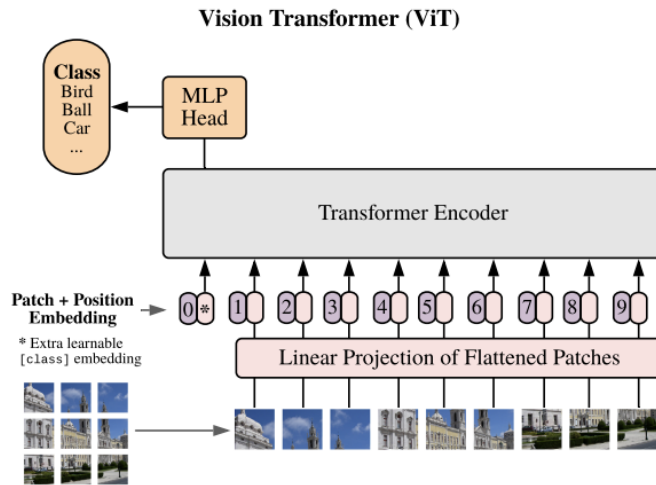
Google Research, Brain Team; <sup>†</sup>independent researcher

{andstein, akolesnikov, xzhai, usz, lbeyer}@google.com, rwightman@gmail.com



### 3. How to train your ViT?

- ❖ How To Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers
  - Models: ViT-Ti, ViT-S, ViT-B , ViT-L
  - Patch-size: 16, 32(ViT-S, ViT-B)
  - 기존 ViT에서 head의 hidden layer를 제거
  - Backbone으로 ResNet을 사용하여 나온 특징벡터를 ViT에 입력(Hybrid models)



<ViT architecture>

Table 1: Configurations of ViT models.

Model	Layers	Width	MLP	Heads	Params
ViT-Ti [34]	12	192	768	3	5.8M
ViT-S [34]	12	384	1536	6	22.2M
ViT-B [10]	12	768	3072	12	86M
ViT-L [10]	24	1024	4096	16	307M

Table 2: ResNet+ViT hybrid models.

Model	Resblocks	Patch-size	Params
R+Ti/16	[]	8	6.4M
R26+S/32	[2, 2, 2, 2]	1	36.6M
R50+L/32	[3, 4, 6, 3]	1	330.0M

# 3. How to train your ViT?

---

## ❖ How To Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers

- Datasets for pre-training
  - ImageNet-1k
  - ImageNet21k
    - De-duplicate images in ImageNet-21k
- Datasets to evaluate transfer learning
  - VTAB benchmark
  - CIFAR-100
  - Oxford IIIT Pets
  - Resisc45 and Kittidistance

### 3. How to train your ViT?

---

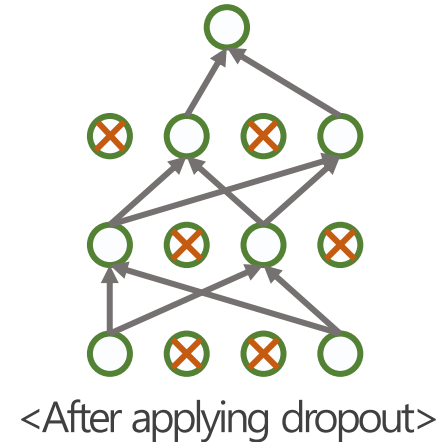
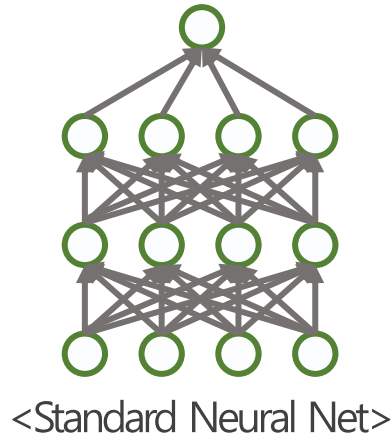
#### ❖ How To Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers

##### ➤ Regularization

- **Dropout to intermediate activations of ViT**
- Stochastic depth regularization technique

##### ➤ Data augmentations

- Mixup
- RandAugment



### 3. How to train your ViT?

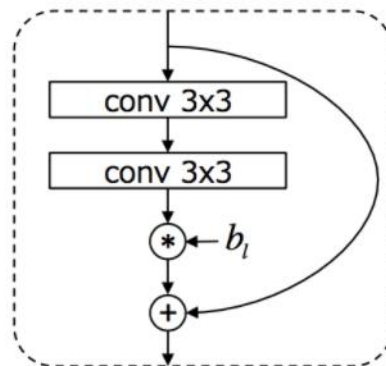
#### ❖ How To Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers

##### ➤ Regularization

- Dropout to intermediate activations of ViT
- **Stochastic depth regularization technique**

##### ➤ Data augmentations

- Mixup
- RandAugment



<|번째 residual block>

$$pl = 1 - \frac{l}{2L}$$

### 3. How to train your ViT?

#### ❖ How To Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers

- Regularization
  - Dropout to intermediate activations of ViT
  - Stochastic depth regularization technique
- **Data augmentations**
  - **Mixup**
  - RandAugment



$[1, 0]$



$\lambda=0.5$



$[0, 1]$

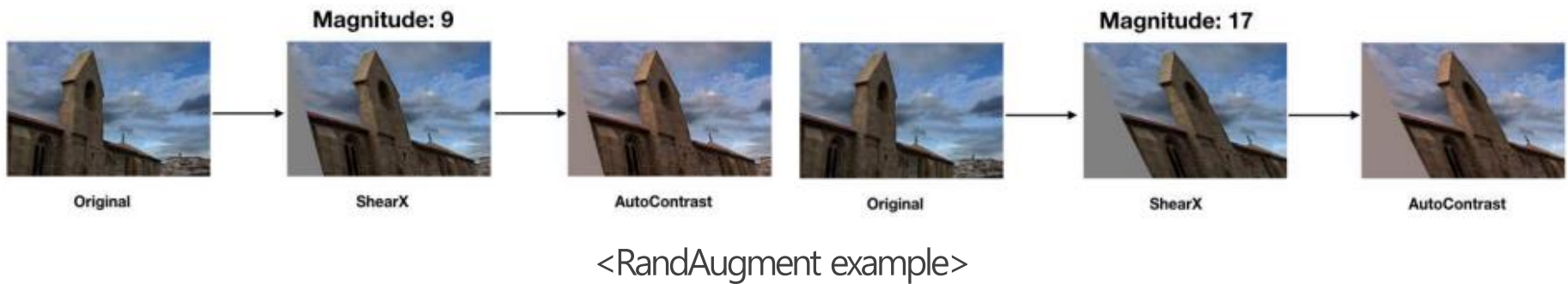


$[0.5, 0.5]$

# 3. How to train your ViT?

## ❖ How To Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers

- Regularization data augmentations
  - Dropout to intermediate activations of ViT
  - Stochastic depth regularization technique
- **Data augmentations**
  - Mixup
  - **RandAugment**



RandAugment: Practical automated data augmentation with a reduced search space

# 3. How to train your ViT?

---

## ❖ How To Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers

### ➤ Pre-training

- Optimizer: Adam
- Batch size: 4096
- Cosine learning rate schedule
- Image pre-process: Inception-style cropping, random horizontal flipping
- Epoch: ImageNet-1k(300), ImageNet-21k(30 & 300)

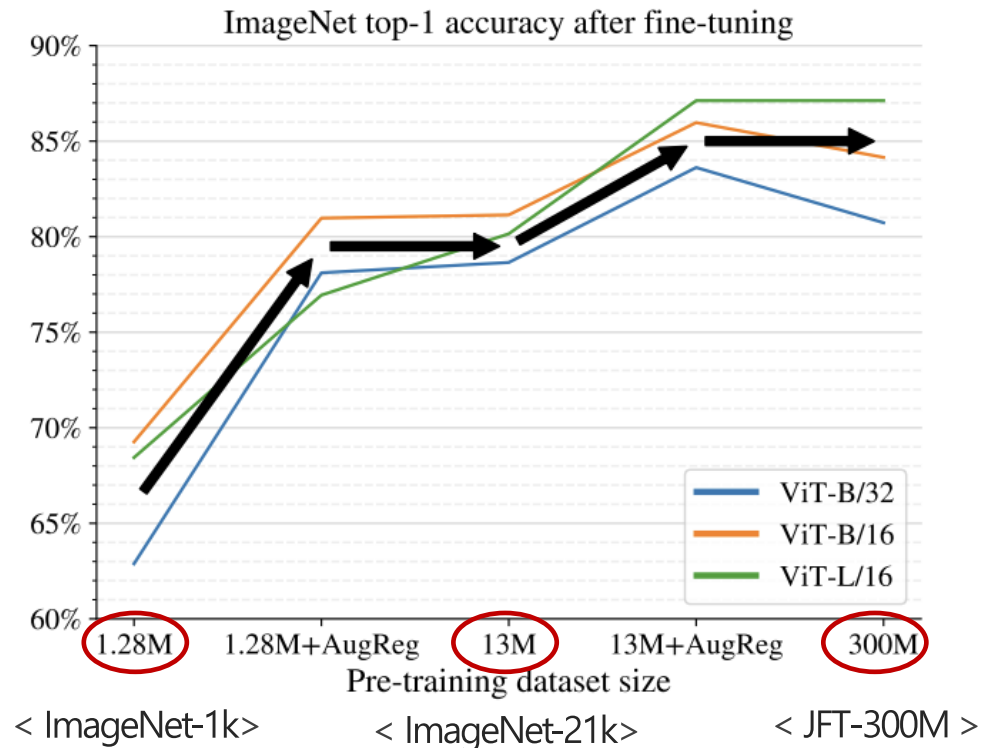
### ➤ Fine-tuning

- Optimizer: SGD
- Batch size: 512
- Cosine decay learning rate schedule

### 3. How to train your ViT?

#### ❖ How To Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers

- Scaling datasets with AugReg(augmentation, Reugularization) and compute
  - AugReg ImageNet-1k perform about equal to the same models pre-trained ImageNet-21k dataset
  - AugReg ImageNet-21k perform about equal to the same models pre-trained plain JFT-300M dataset

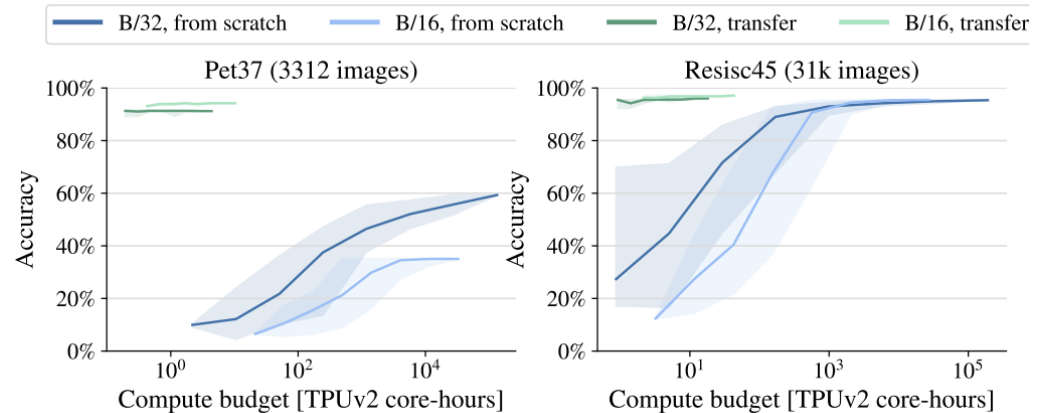
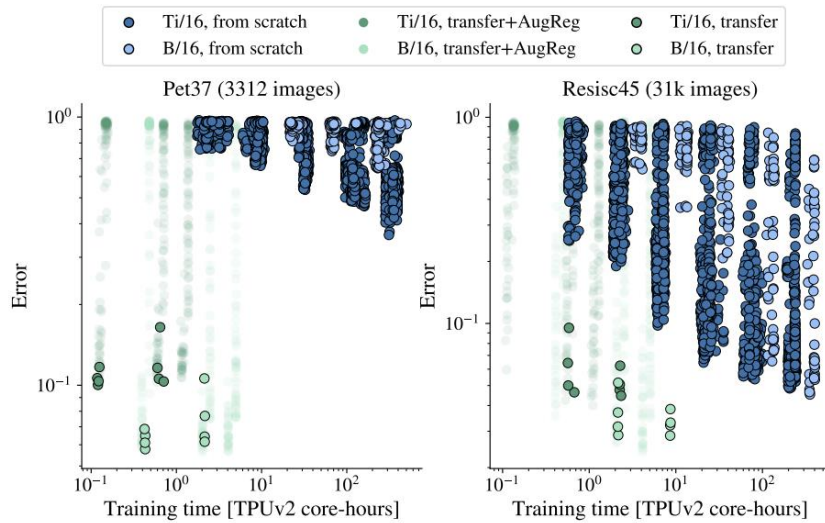




# 3. How to train your ViT?

## ❖ How To Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers

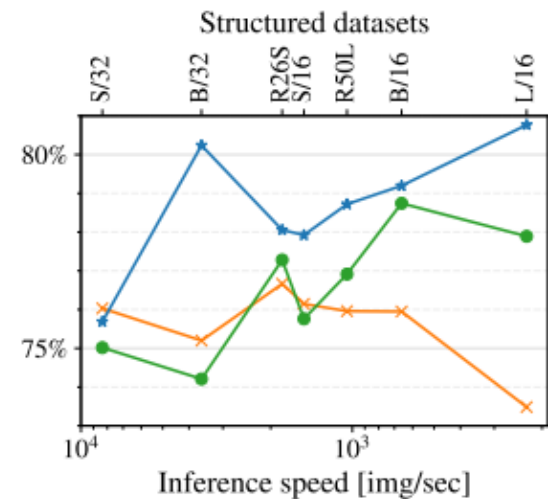
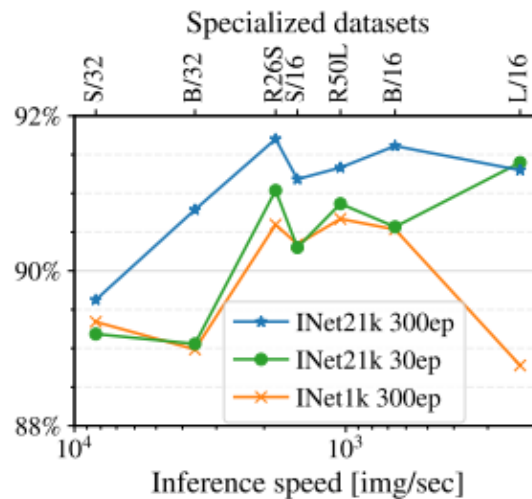
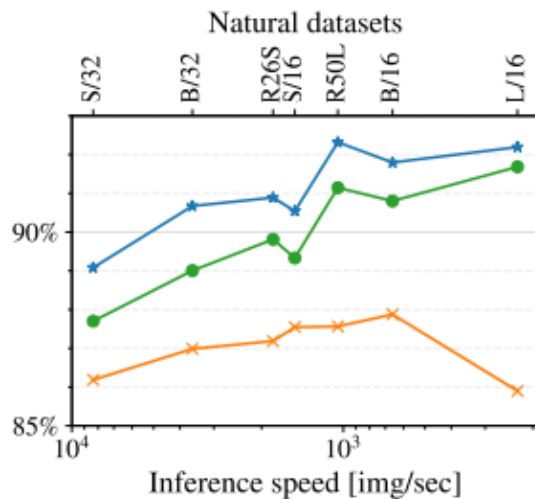
- Transfer is the better option
  - Transferring a pre-trained model is both more cost-efficient and leads to better results
  - It does not seem possible to train ViT models from scratch to reach accuracy anywhere near that of transferred models
  - The ViT models from scratch are impossible to train with the accuracy of transferred models



# 3. How to train your ViT?

## ❖ How To Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers

- More data yields more generic models
  - Model pre-trained on ImageNet-21k(30 epochs) is better than the ImageNet-1k(300 epochs)
  - **More data yields more generic models**



### 3. How to train your ViT?

#### ❖ How To Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers

- Prefer augmentation to regularization
  - Mid-sized ImageNet-1k dataset, any kind of AugReg helps
  - ImageNet-21k dataset and keeping compute fixed, any kind of AugReg hurts performance for all but the largest models

ImageNet-1k, 300ep															ImageNet-21k, 30ep															ImageNet-21k, 300ep															
No regularization										Regularization 0.1					No regularization										Regularization 0.1					No regularization										Regularization 0.1					
RTi	69	73	73	72	70	69	68	71	70	67	65	63	62	61	-2	36	35	33	32	31	30	29	33	31	28	27	26	25	24	-3	39	38	37	36	35	34	33	36	34	32	31	30	29	28	-3
Ti/16	72	76	75	75	74	72	71	71	72	68	65	63	63	62	-4	38	37	35	34	33	32	31	34	32	29	27	26	25	25	-4	41	41	40	39	37	37	36	38	36	34	33	32	31	29	-4
S/32	64	71	76	76	76	74	74	70	72	72	71	71	69	68	-4	40	39	38	37	35	35	34	37	34	33	32	31	30	29	-3	43	43	43	42	42	41	40	41	40	39	38	36	36	35	-2
S/16	71	77	79	81	82	80	80	76	79	80	79	79	77	77	-2	44	43	43	42	41	40	39	42	40	39	38	37	35	35	-2	46	47	47	47	46	45	45	45	45	43	43	42	41	40	-2
B/32	63	70	73	75	76	75	76	69	74	77	77	78	77	77	2	43	42	43	42	41	40	40	42	40	40	38	38	36	36	-1	42	46	48	47	47	47	46	45	46	46	45	44	44	43	-2
R26S	72	76	78	79	80	80	80	75	78	81	82	82	81	81	2	45	45	43	43	42	41	41	44	44	42	41	40	40	39	-1	46	48	48	47	47	46	46	48	48	46	46	45	44	43	-0
B/16	70	76	79	79	81	80	80	76	79	81	82	83	82	82	2	46	47	47	46	46	45	44	46	45	45	44	43	42	41	-1	43	48	50	51	50	50	50	47	49	49	49	48	48	48	-1
L/16	69	76	77	78	78	76	76	74	78	78	78	79	77	77	0	45	48	49	49	49	48	47	48	48	48	47	46	45	45	-1	43	46	49	49	51	51	51	46	48	51	51	51	51	51	-0
R50L	70	75	76	77	77	76	76	75	78	78	78	79	77	77	2	45	47	48	48	48	47	47	48	48	48	47	47	45	46	0	42	46	47	49	51	50	51	46	48	51	51	51	51	51	1
	none	light1	light2	med1	med2	heavy1	heavy2	none	light1	light2	med1	med2	heavy1	heavy2	reg - noreg	none	light1	light2	med1	med2	heavy1	heavy2	none	light1	light2	med1	med2	heavy1	heavy2	reg - noreg	none	light1	light2	med1	med2	heavy1	heavy2	none	light1	light2	med1	med2	heavy1	heavy2	reg - noreg

### 3. How to train your ViT?

#### ❖ How To Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers

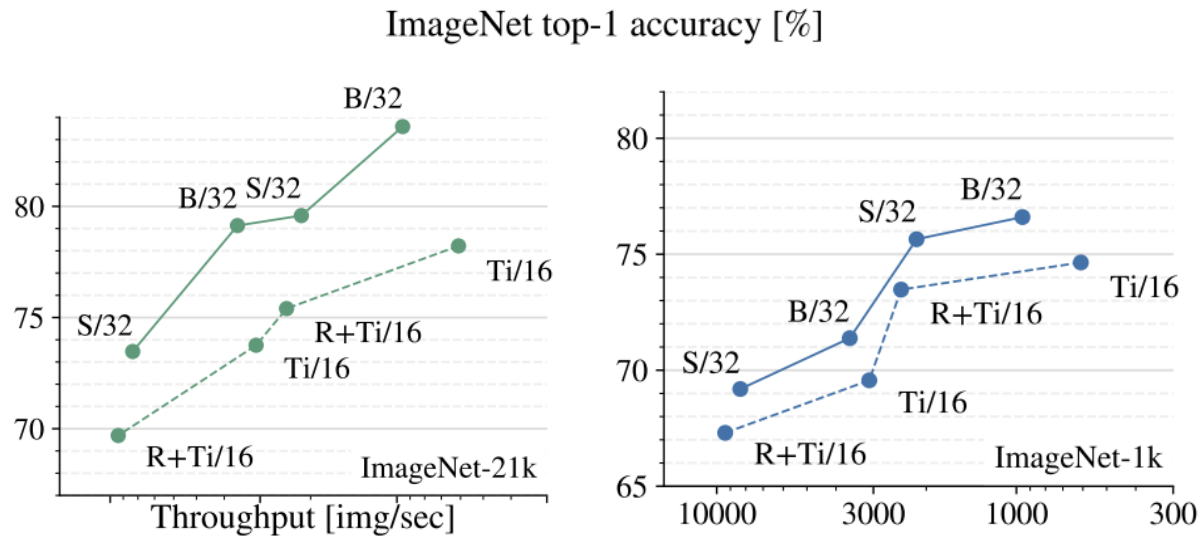
- Choosing which pre-trained model to transfer
  - **How to select a model for further adaption for an end application?**
  - Run downstream adaptation for all available pre-trained models
  - One can select a single pre-trained model based on the upstream validation accuracy, which is much cheaper
  - Cheaper strategy works equally well as the more expensive strategy in the majority of scenarios

	INet-1k (v2 val)	CIFAR-100	Pets	Resisc45	Kitti	ImageNet-1k
RTi	-0.7	+0.4	+0.5	-1.5	+0.1	+0.0
Ti/16	-0.7	+0.0	+0.6	+2.8	+0.4	+0.0
S/32	-1.3	+1.3	-0.0	+2.1	-0.4	+0.2
S/16	+0.1	-0.1	-0.5	+0.3	+0.6	+1.5
B/32	+0.2	+0.0	+0.0	+0.0	+0.0	+6.2
R26S	+0.0	-0.2	-0.2	+0.8	+0.0	+0.8
B/16	+0.2	-0.0	+0.3	-3.6	-0.7	+3.9
L/16	-0.3	+0.5	-0.3	+1.3	-1.5	+1.0
R50L	-0.3	+0.6	+0.3	+0.6	-0.2	+3.7

### 3. How to train your ViT?

#### ❖ How To Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers

- Prefer increasing patch-size to shrinking model-size
  - “Tiny” variants perform significantly worse than the similarly fast larger models with “/32” patch-size
  - **Parameter count is reflective neither of speed, nor of capacity**



## 4. Conclusion

---

### ❖ 결론

#### ➤ ViT

- Computer vision분야에서 합성곱 신경망이 아닌 Transformer를 적용
- Transformer architecture가 computer vision분야에서 수 많은 SOTA를 달성

#### ➤ How To Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers

- Regularization과 augmentation만 잘 써도 적은 데이터로 비슷한 성능을 도출 할 수 있음
- Downstream task시 대용량 데이터로 사전 학습한 모델로 transfer learning하는 것이 좋음
- Transfer learning시 대용량 데이터로 사전학습한 모델의 성능이 좋음
- 데이터가 많이 존재할 경우 regularization과 augmentation은 효과가 거의 없음
- 사전학습 모델이 많은 경우 upstream에서 가장 성능 좋은 모델을 사용하면 좋음
- Inference의 속도가 중요한 경우 patch size를 키우면 성능이 유지됨

---

# 감사합니다.