DeiT

Data-Efficient Image Transformers

Segmentation – 16기 분석 이지혜

01. Introduction

02. Prerequisites
- Vision Transformer
- Knowledge Distillation

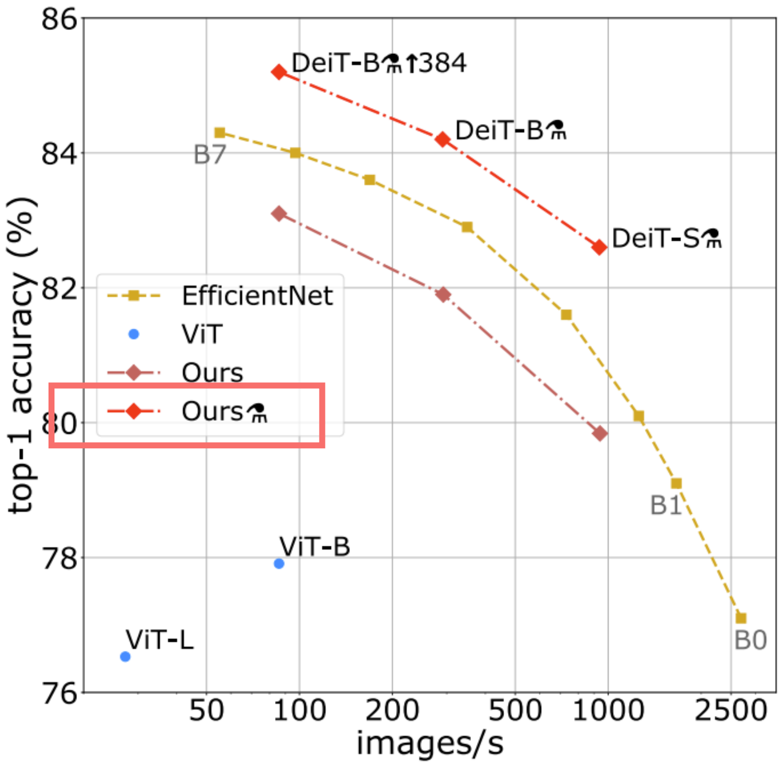
03. Architecture
- Bag of Tricks
- Knowledge Distillation

04. Experiments
- Transformer Models
- Distillation
- Efficiency vs Accuracy
- Transfer Learning

PART 1

Introduction

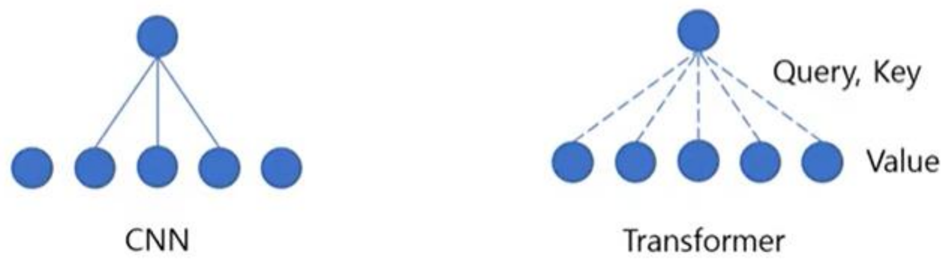
- High Performing Vision Transformers on image understanding tasks using large infrastructure -> LIMITS
- Convolution-Free Transformers
- Teacher-Student Strategies
- Token-Based Distillation

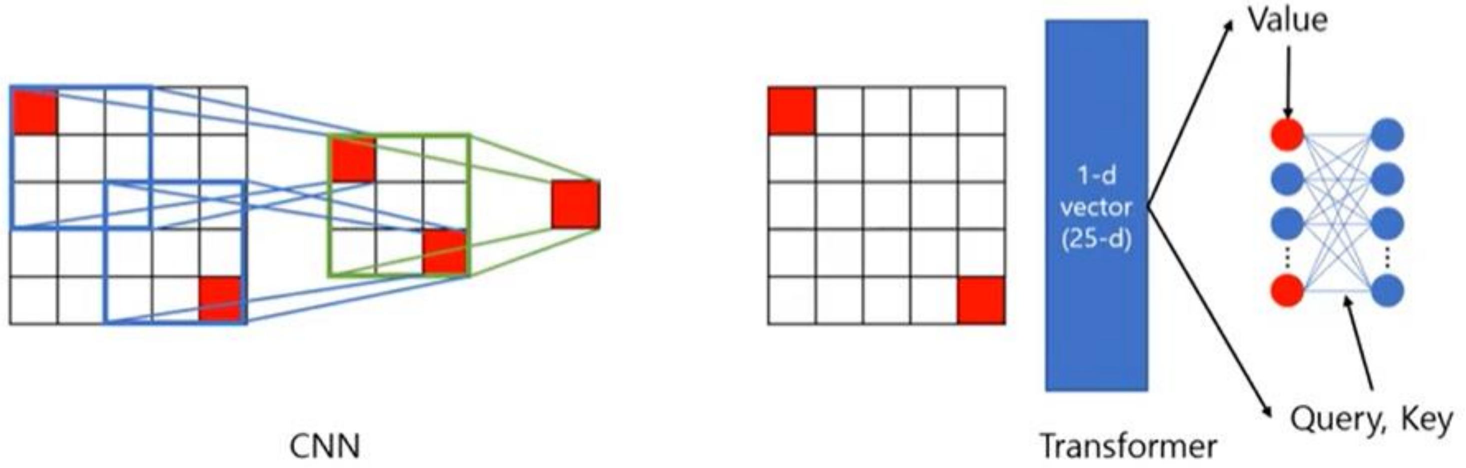


Transformer vs CNN

- **CNN** : 이미지 전체의 정보를 취합하기 위해 몇 개의 **layer** 통과

- **Transformer** : 하나의 **layer**만으로 전체 이미지 정보 취합 가능





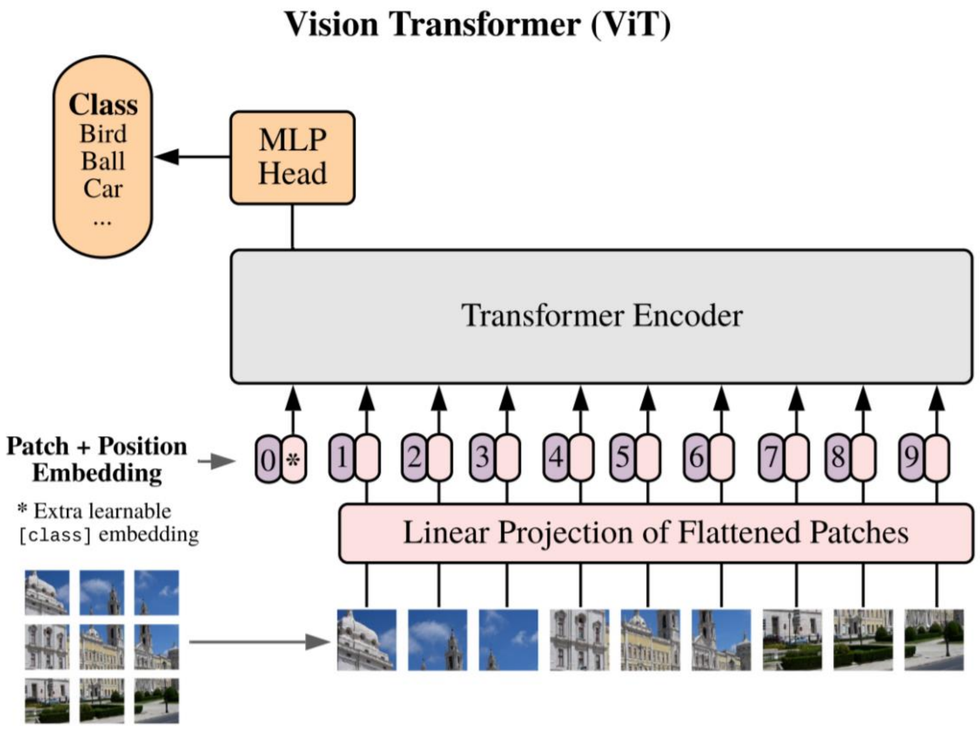
Prerequisites

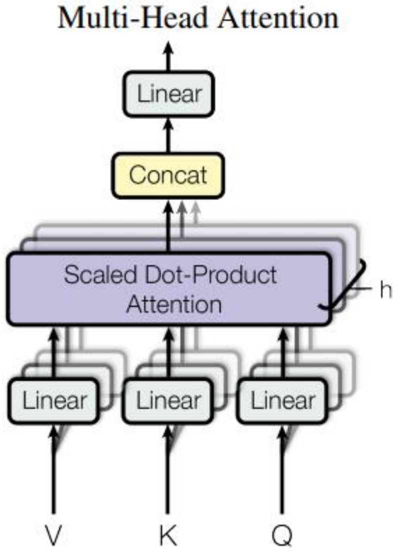
Vision Transformer

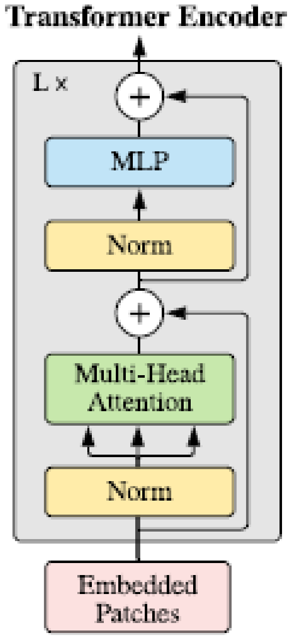
- Training Dataset : JFT-300M
- Pre-Train : Low Resolution
- Fine-Turning : High Resolution
- Position Embedding : Bicubic Interpolation

	Ours (ViT-H/14)	Ours (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	88.36	87.61 ± 0.03	87.54 ± 0.02	88.4/**88.5***
ImageNet ReaL	**90.77**	90.24 ± 0.03	90.54	90.55
CIFAR-10	$\mathbf{99.50} \pm 0.06$	99.42 ± 0.03	99.37 ± 0.06	–
CIFAR-100	$\mathbf{94.55} \pm 0.04$	93.90 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	$\mathbf{97.56} \pm 0.03$	97.32 ± 0.11	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	$\mathbf{99.74} \pm 0.00$	99.63 ± 0.03	–
VTAB (19 tasks)	$\mathbf{77.16} \pm 0.29$	75.91 ± 0.18	76.29 ± 1.70	–
TPUv3-days	2.5k	0.68k	9.9k	12.3k

Vision Transformer(VIT)







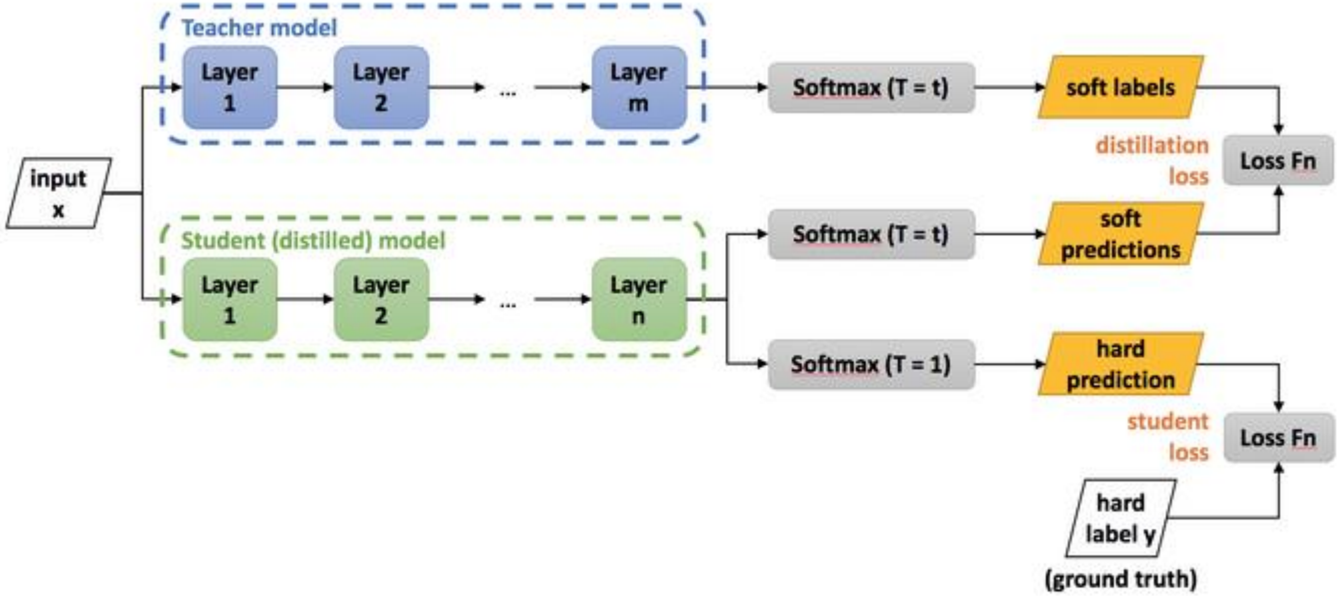
- Patch Embedding
- Resolution : (H, W) -> (P, P)
- 2D Interpolation of pre-trained Position Embedding

- Multi-Head Self Attention Layers
- Transformer Block for Image
- Fixing the Positional Encoding across Resolutions

Prerequisites

Knowledge Distillation

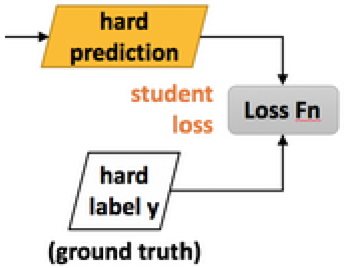
- A student model learns from a larger teacher model



The Loss Function

$$Total\ Loss = (1-\alpha)L_{CE}(\sigma(Z_s),\ \hat{y}) + 2\alpha T^2 L_{CE}(\sigma(\frac{Z_s}{T}),\sigma(\frac{Z_s}{T}))$$

- Student Loss



$$(1-\alpha)L_{CE}(\sigma(Z_s),\ \hat{y})$$

$L_{CE}()$: *Cross entropy loss*

$\sigma()$: *Softmax*

Z_s: *Output logits of Student network*

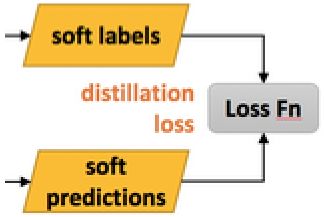
Z_t: *Output logits of Teacher network*

$\hat{y}$: *Ground truth(one-hot)*

α: *Balancing parameter*

T: *Temperature hyperparameter*

- Distillation Loss



$$2\alpha T^2 L_{CE}(\sigma(\frac{Z_s}{T}),\sigma(\frac{Z_s}{T}))$$

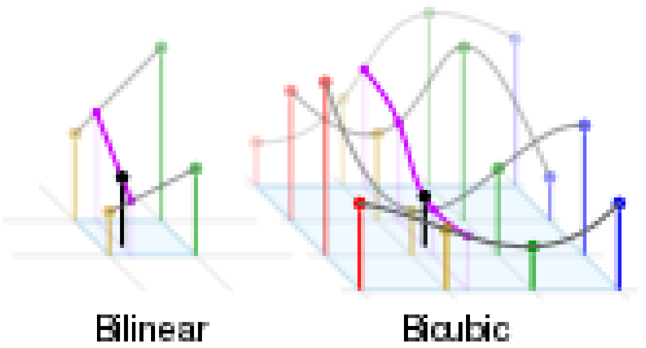
Architecture

Bag of Tricks

- Using the Architecture of VIT (VIT-B = DeiT-B)
- Training Method same as VIT
- Added Hyper Parameter Tuning

PART3-1

Hyper Parameter Tuning



Ablation on ↓	Pre-training	Fine-tuning	Rand-Augment	AutoAug	Mixup	CutMix	Erasing	Stoch. Depth	Repeated Aug.	Dropout	Exp. Moving Avg.	top-1 accuracy pre-trained 224^2	top-1 accuracy fine-tuned 384^2
none: DeiT-B	adamw	adamw	✓	✗	✓	✓	✓	✓	✓	✗	✗	81.8 ±0.	83.1 ±0.1
optimizer	SGD	adamw	✓	✗	✓	✓	✓	✓	✓	✗	✗	74.5	77.3
	adamw	SGD	✓	✗	✓	✓	✓	✓	✓	✗	✗	81.8	83.1
data augmentation	adamw	adamw	✗	✗	✓	✓	✓	✓	✓	✗	✗	79.6	80.4
	adamw	adamw	✗	✓	✓	✓	✓	✓	✓	✗	✗	81.2	81.9
	adamw	adamw	✓	✗	✗	✓	✓	✓	✓	✗	✗	78.7	79.8
	adamw	adamw	✓	✗	✓	✗	✓	✓	✓	✗	✗	80.0	80.6
	adamw	adamw	✓	✗	✗	✗	✓	✓	✓	✗	✗	75.8	76.7
regularization	adamw	adamw	✓	✗	✓	✓	✗	✓	✓	✗	✗	4.3*	0.1
	adamw	adamw	✓	✗	✓	✓	✓	✗	✓	✗	✗	3.4*	0.1
	adamw	adamw	✓	✗	✓	✓	✓	✓	✗	✗	✗	76.5	77.4
	adamw	adamw	✓	✗	✓	✓	✓	✓	✓	✓	✗	81.3	83.1
	adamw	adamw	✓	✗	✓	✓	✓	✓	✓	✗	✓	81.9	83.1

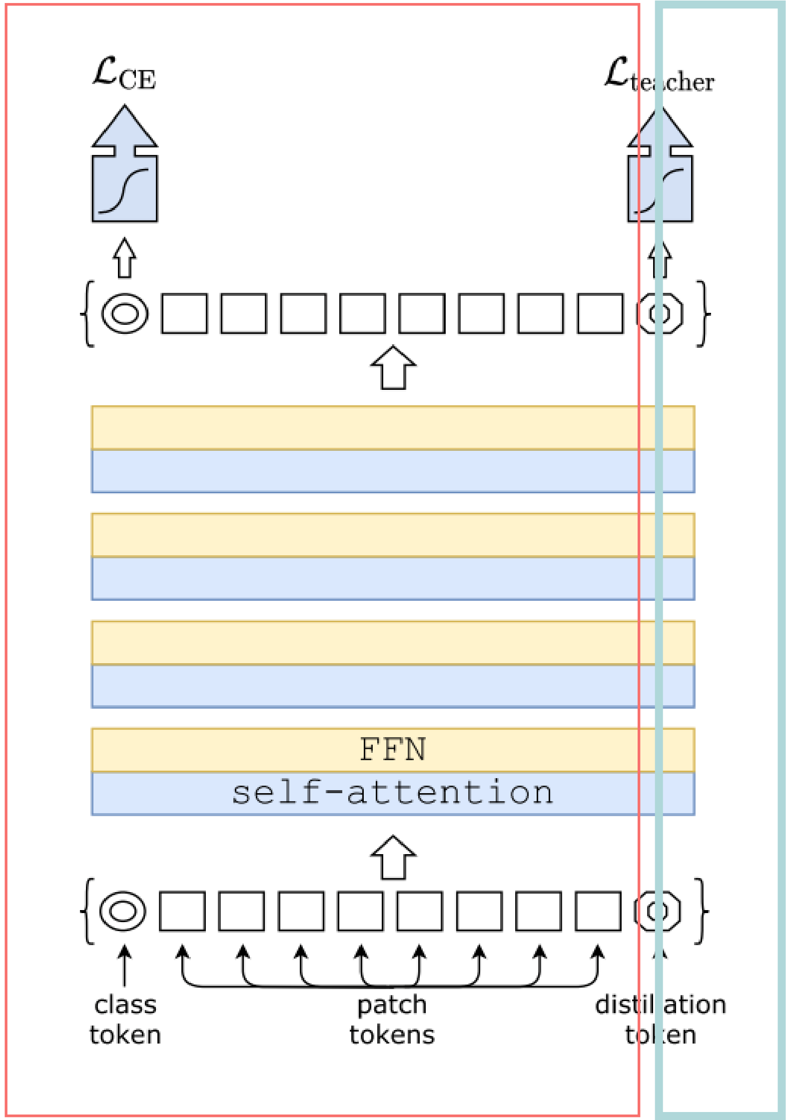
Methods	ViT-B [15]	DeiT-B
Epochs	300	300
Batch size	4096	1024
Optimizer	AdamW	AdamW
learning rate	0.003	$0.0005 \times \frac{\text{batchsize}}{512}$
Learning rate decay	cosine	cosine
Weight decay	0.3	0.05
Warmup epochs	3.4	5
Label smoothing ε	✗	0.1
Dropout	0.1	✗
Stoch. Depth	✗	0.1
Repeated Aug	✗	✓
Gradient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

Architecture

Knowledge Distillation

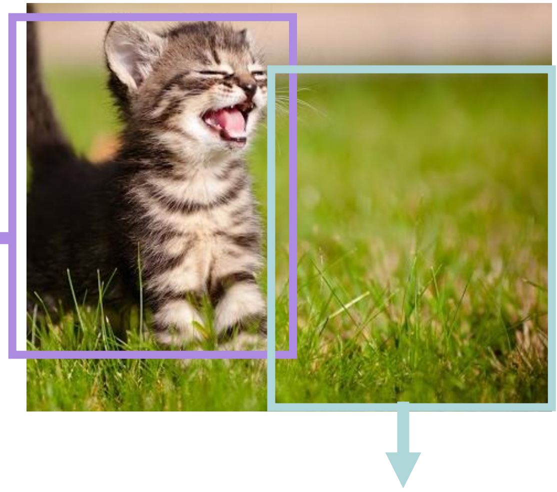
- Adding Distillation Token
- Joint Classifiers
- Better with ConvNet as Teacher Network

Tokenized Distillation



- Simply Include a New Distillation Token
- Interacts with the Class, Path Tokens
- Network's objective is to reproduce the Hard Label Predicted by the Teacher Network

Soft Distillation vs Hard Distillation



- Hard Distillation

$$\mathcal{L}_{\text{global}}^{\text{hardDistill}} = \frac{1}{2}\mathcal{L}_{\text{CE}}(\psi(Z_s), y) + \frac{1}{2}\mathcal{L}_{\text{CE}}(\psi(Z_s), y_{\text{t}}).$$

- Soft Distillation

$$\mathcal{L}_{\text{global}} = (1-\lambda)\mathcal{L}_{\text{CE}}(\psi(Z_{\text{s}}), y) + \lambda\tau^2 \text{KL}(\psi(Z_{\text{s}}/\tau), \psi(Z_{\text{t}}/\tau)).$$

Distillation

- Teacher Model : RegNetY-16GF
- Inductive Bias

Teacher Models	acc.	Student: DeiT-B ⚗ pretrain	↑384
DeiT-B	81.8	81.9	83.1
RegNetY-4GF	80.0	82.7	83.6
RegNetY-8GF	81.7	82.7	83.8
RegNetY-12GF	82.4	83.1	84.1
RegNetY-16GF	82.9	83.1	84.2

- Learns Better from Distillation Method of Convnet

	groundtruth	no distillation convnet	DeiT	DeiT⚗ student (of the convnet) class	distillation	DeiT⚗
groundtruth	0.000	0.171	0.182	0.170	0.169	0.166
convnet (RegNetY)	0.171	0.000	0.133	0.112	0.100	0.102
DeiT	0.182	0.133	0.000	0.109	0.110	0.107
DeiT⚗– class only	0.170	0.112	0.109	0.000	0.050	0.033
DeiT⚗– distil. only	0.169	0.100	0.110	0.050	0.000	0.019
DeiT⚗– class+distil.	0.166	0.102	0.107	0.033	0.019	0.000

- Distillation Comparison : Hard Distillation Is Better

method ↓	Supervision label	teacher	ImageNet top-1 (%) Ti 224	S 224	B 224	B↑384
DeiT– no distillation	✓	✗	72.2	79.8	81.8	83.1
DeiT– usual distillation	✗	soft	72.2	79.8	81.8	83.2
DeiT– hard distillation	✗	hard	74.3	80.9	83.0	84.0
DeiT⚗: class embedding	✓	hard	73.9	80.9	83.0	84.2
DeiT⚗: distil. embedding	✓	hard	74.6	81.1	83.1	84.4
DeiT⚗: class+distillation	✓	hard	74.5	81.2	83.4	84.5

Efficiency vs Accuracy

Model	ViT model	embedding dimension	#heads	#layers	#params	training resolution	throughput (im/sec)
DeiT-Ti	N/A	192	3	12	5M	224	2536
DeiT-S	N/A	384	6	12	22M	224	940
DeiT-B	ViT-B	768	12	12	86M	224	292

Model	#Param	Image throughput	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_dist 384	87M	85.8	84.5

Data Augmentation

Knowledge Distillation

Network	#param.	image size	throughput (image/s)	ImNet top-1	Real top-1	V2 top-1
Convnets						
ResNet-18 [21]	12M	224^2	4458.4	69.8	77.3	57.1
ResNet-50 [21]	25M	224^2	1226.1	76.2	82.5	63.3
ResNet-101 [21]	45M	224^2	753.6	77.4	83.7	65.7
ResNet-152 [21]	60M	224^2	526.4	78.3	84.1	67.0
RegNetY-4GF [40]⋆	21M	224^2	1156.7	80.0	86.4	69.4
RegNetY-8GF [40]⋆	39M	224^2	591.6	81.7	87.4	70.8
RegNetY-16GF [40]⋆	84M	224^2	334.7	82.9	88.1	72.4
EfficientNet-B0 [48]	5M	224^2	2694.3	77.1	83.5	64.3
EfficientNet-B1 [48]	8M	240^2	1662.5	79.1	84.9	66.9
EfficientNet-B2 [48]	9M	260^2	1255.7	80.1	85.9	68.8
EfficientNet-B3 [48]	12M	300^2	732.1	81.6	86.8	70.6
EfficientNet-B4 [48]	19M	380^2	349.4	82.9	88.0	72.3
EfficientNet-B5 [48]	30M	456^2	169.1	83.6	88.3	73.6
EfficientNet-B6 [48]	43M	528^2	96.9	84.0	88.8	73.9
EfficientNet-B7 [48]	66M	600^2	55.1	84.3	_	_
EfficientNet-B5 RA [12]	30M	456^2	96.9	83.7	_	_
EfficientNet-B7 RA [12]	66M	600^2	55.1	84.7	_	_
KDforAA-B8	87M	800^2	25.2	85.8	_	_
Transformers						
ViT-B/16 [15]	86M	384^2	85.9	77.9	83.6	_
ViT-L/16 [15]	307M	384^2	27.3	76.5	82.2	_
DeiT-Ti	5M	224^2	2536.5	72.2	80.1	60.4
DeiT-S	22M	224^2	940.4	79.8	85.7	68.5
DeiT-B	86M	224^2	292.3	81.8	86.7	71.5
DeiT-B↑384	86M	384^2	85.9	83.1	87.7	72.4
DeiT-Ti⚗	6M	224^2	2529.5	74.5	82.1	62.9
DeiT-S⚗	22M	224^2	936.2	81.2	86.8	70.0
DeiT-B⚗	87M	224^2	290.9	83.4	88.3	73.2
DeiT-Ti⚗ / 1000 epochs	6M	224^2	2529.5	76.6	83.9	65.4
DeiT-S⚗ / 1000 epochs	22M	224^2	936.2	82.6	87.8	71.7
DeiT-B⚗ / 1000 epochs	87M	224^2	290.9	84.2	88.7	73.9
DeiT-B⚗ ↑384	87M	384^2	85.8	84.5	89.0	74.8
DeiT-B⚗ ↑384 / 1000 epochs	87M	384^2	85.8	85.2	89.3	75.2

감사합니다