Training data-efficient image transformers & distillation through attention

School of Industrial and Management Engineering, Korea University

Eun Ji Koh

KOREA UNIVERSITY

DMQA

Contents

Research Purpose

❖ Training data-efficient image transformers & distillation through attention (arXiv, 2020)

- Facebook AI, Sorbonne University에서 연구하였으며 2021년 07월 12일 기준으로 186회 인용됨

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.

Research Purpose

❖ Training data-efficient image transformers & distillation through attention (arXiv, 2020)

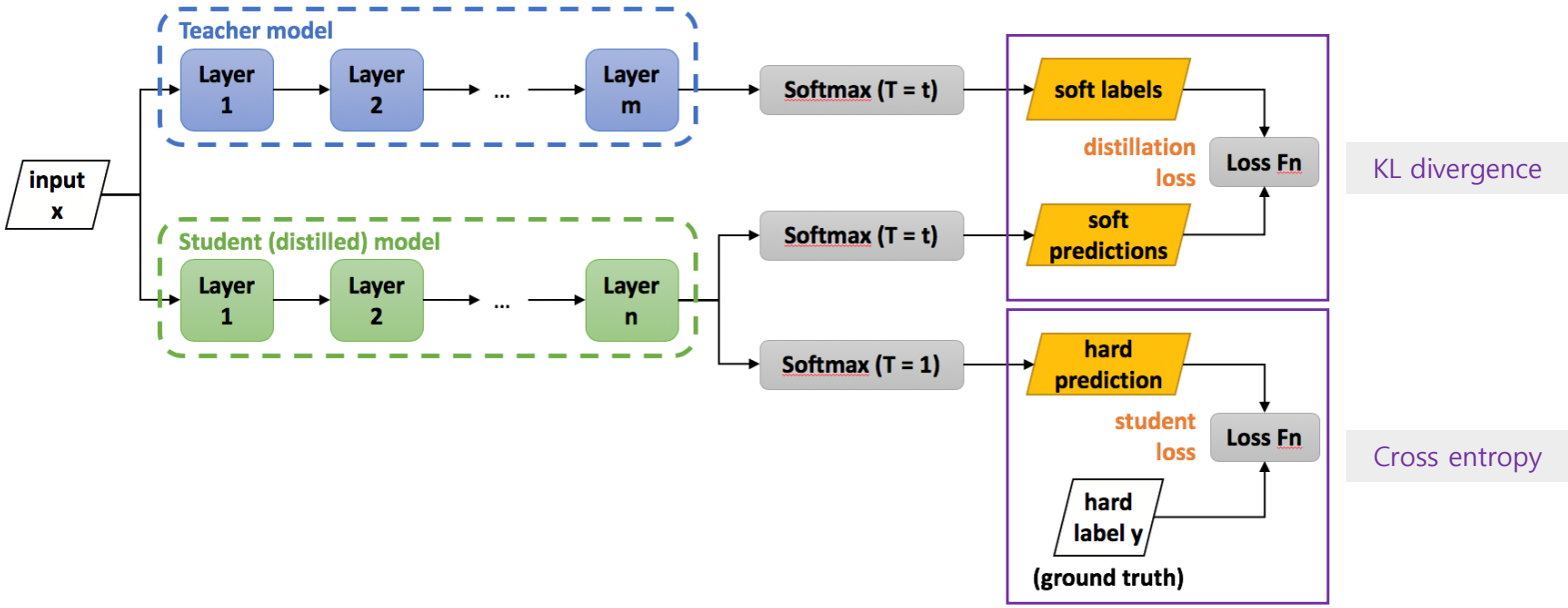
- ViT의 한계점
  - ➢ Transformer 사용으로 인해 inductive bias를 학습하지 못하기 때문에, 다량의 학습 데이터를 사용하지 않는 경우 generalize 및 SOTA 달성에 한계가 있음
  - ➢ extensive한 computing resources 필요

- Knowledge Distillation 기법을 적용하여 대규모의 데이터 셋을 활용한 pre-training 과정 없이 ImageNet 데이터 셋만 사용하여 좋은 성능을 내고자 함
  - ➢ Attention을 통해 student 모델이 teacher 모델로부터 학습할 수 있게 하는 Distillation token을 활용한 teacher-student strategy 제안

Data-efficient image Transformers (Deit)

❖ Diagram of DeiT (Knowledge Distillation)

- 잘 학습된 teacher model의 지식을 student model로 전이하여 상대적으로 작은 student model도 큰 teacher model에 필적하는 성능을 낼 수 있도록 함
- Teacher의 모델의 softmax 분포와 student 모델의 softmax 분포의 KL divergence를 최소화



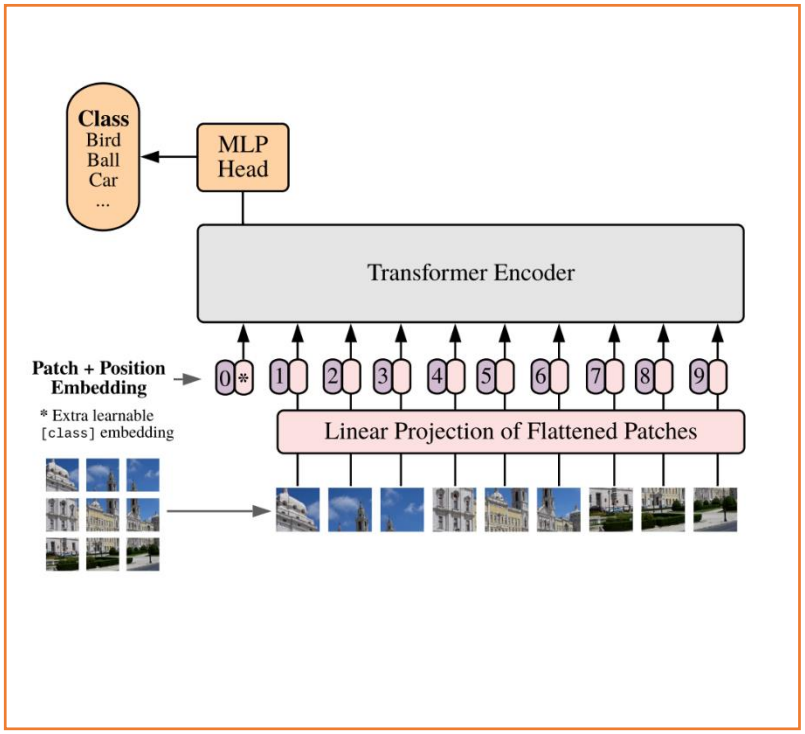
https://intellabs.github.io/distiller/knowledge_distillation.html

Data-efficient image Transformers (Deit)

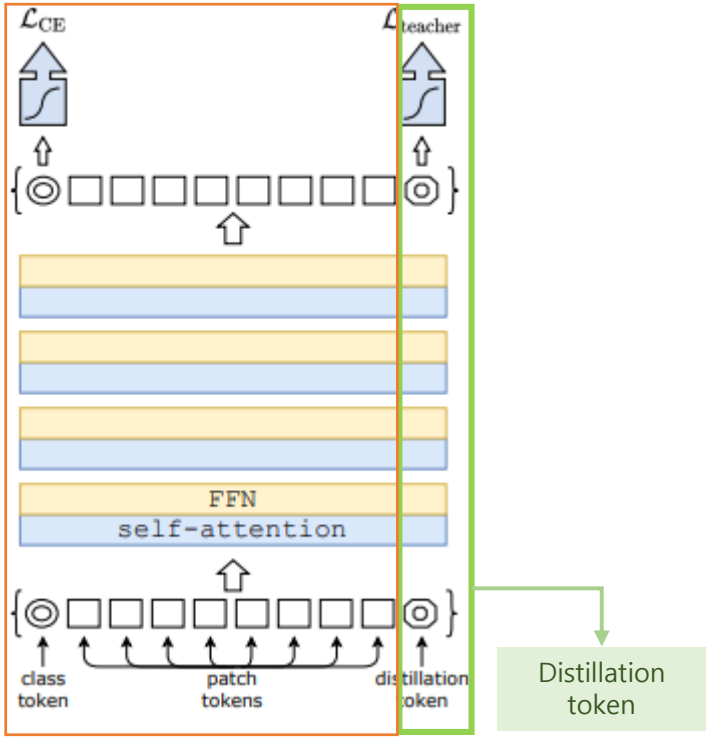
❖ Diagram of DeiT (Distillation token)

- ViT와 동일한 아키텍처에 distillation token을 추가
- Distillation token은 class token의 학습 방식과 유사하게, teacher model output과의 cross entropy를 통해 학습

Vision Transformer (ViT)



Data-efficient image transformers (DeiT)



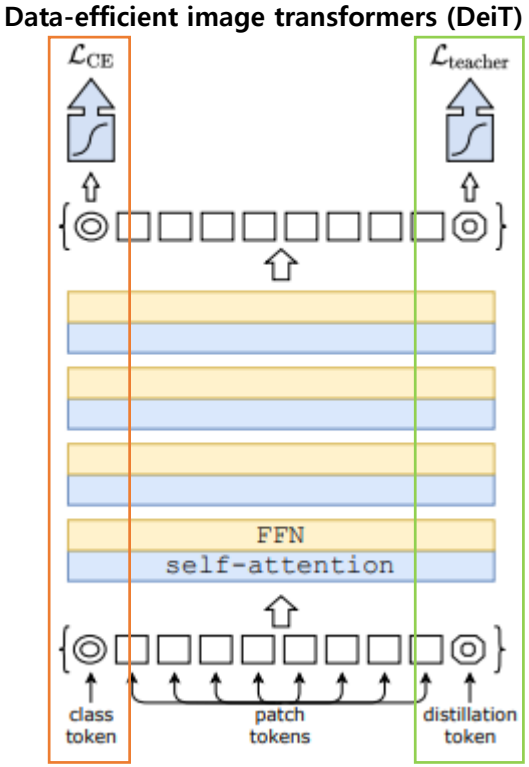
Data-efficient image Transformers (Deit)

❖ Diagram of DeiT (Distillation token)

- Class token과 distillation token의 코사인 유사도

  : 실험 시, 초기에는 0.06이었으나 마지막 embedding 상태에서는 0.93 (1보다 작은 값)

- 따라서 class token과 distillation token의 output이 같은 것은 아님 (단순히 유사한 output일 뿐)



Experiments

❖ DeiT model size와 teacher model에 따른 비교 실험

- Base model (DeiT-B)의 parameter 수가 가장 크고, interpolate방식으로 파라미터 값들을 줄임
- Teacher model로 DeiT-B를 사용하는 것보다 CNN계열 모델을 쓰는 경우의 accuracy가 더 높음
  - ➢ Distillation을 통해 CNN계열 모델의 inductive bias를 학습 가능하기 때문

	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
Base model	DeiT-B	ViT-B	768	12	12	86M	224	292

- DeiT-B : Base model. Same as ViT-B
- DeiT-B↑384: DeiT-B를 더 높은 해상도의 이미지로 학습시킨 model
- DeiT ⚗ : DeiT with distillation (distillation token)
- DeiT-S (Small), DeiT-Ti (Tiny): DeiT의 경량 모델

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

Experiments

❖ Distillation 방식 및 token 사용 여부에 따른 비교 실험

- Hard distillation를 사용하는 경우의 accuracy가 대체로 높음
- 추가적으로 distillation token 및 class token을 동시에 사용할 때 accuracy 가장 높음

Distillation 방식

Distillation token 사용 여부

method ↓	Supervision label	Supervision 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

Experiments

❖ Distillation 방식 및 token 사용 여부에 따른 비교 실험

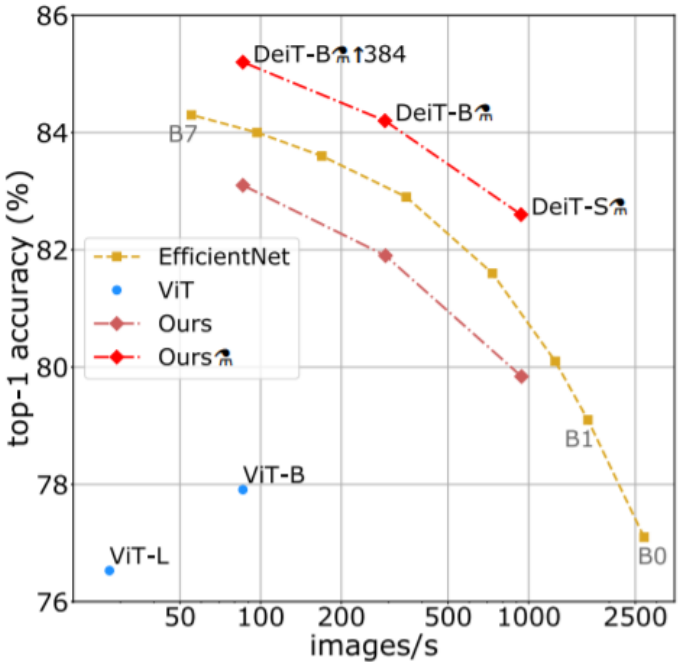
- Inductive bias가 잘 전달되었는지 간접적으로 확인해보고자 함
- 모델 간의 예측이 얼마나 차이 나는지 실험
- Distillation token을 사용한 모델은 DeiT보다 convnet과 상대적으로 유사한 예측을 함

	groundtruth	no distillation		DeiT⚗ student (of the convnet)		
		convnet	DeiT	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

Experiments

❖ Efficiency VS Accuracy

- ImageNet만을 사용하여 학습하면, ViT에 비해 DeiT의 성능이 확실히 좋음



Experiments

❖ Efficiency VS Accuracy

- 다양한 parameter 조합의 모델에 대해 실험
- EfficientNet과 DeiT의 parameter 수가 유사한 경우, DeiT의 속도가 더 빠름
  - ➢ DeiT는 conv filter를 사용하지 않기 때문

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	-	-

Network	#param.	image size	throughput (image/s)	ImNet top-1	Real top-1	V2 top-1
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

Experiments

❖ Transfer Learning에 대한 실험

- 다양한 Dataset을 사용하여 model의 성능 비교
- ViT에 비해 전반적으로 DeiT의 성능이 높게 나타남

Dataset	Train size	Test size	#classes
ImageNet [42]	1,281,167	50,000	1000
iNaturalist 2018 [26]	437,513	24,426	8,142
iNaturalist 2019 [27]	265,240	3,003	1,010
Flowers-102 [38]	2,040	6,149	102
Stanford Cars [30]	8,144	8,041	196
CIFAR-100 [31]	50,000	10,000	100
CIFAR-10 [31]	50,000	10,000	10

Model	ImageNet	CIFAR-10	CIFAR-100	Flowers	Cars	iNat-18	iNat-19	im/sec
Grafit ResNet-50 [49]	79.6	-	-	98.2	92.5	69.8	75.9	1226.1
Grafit RegNetY-8GF [49]	-	-	-	99.0	94.0	76.8	80.0	591.6
ResNet-152 [10]	-	-	-	-	-	69.1	-	526.3
EfficientNet-B7 [48]	84.3	98.9	91.7	98.8	94.7	-	-	55.1
ViT-B/32 [15]	73.4	97.8	86.3	85.4	-	-	-	394.5
ViT-B/16 [15]	77.9	98.1	87.1	89.5	-	-	-	85.9
ViT-L/32 [15]	71.2	97.9	87.1	86.4	-	-	-	124.1
ViT-L/16 [15]	76.5	97.9	86.4	89.7	-	-	-	27.3
DeiT-B	81.8	99.1	90.8	98.4	92.1	73.2	77.7	292.3
DeiT-B↑384	83.1	99.1	90.8	98.5	93.3	79.5	81.4	85.9
DeiT-B⚗	83.4	99.1	91.3	98.8	92.9	73.7	78.4	290.9
DeiT-B⚗↑384	84.4	99.2	91.4	98.9	93.9	80.1	83.0	85.9

Experiments

❖ Training details and Ablation

- Transformer는 상대적으로 initialization에 민감
  - ➢ 아래 논문을 참고하여 hyper parameter 설정 및 실험 진행
  - ➢ Boris Hanin and David Rolnick. How to start training: The effect of initialization and architecture. NIPS, 31, 2018

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

Experiments

❖ Training details and Ablation

- Data Augmentation
  - ➢ Image가 많을 수록 좋은 성능을 보이기 때문에 Extensive 한 방식의 augmentation을 사용
  - ➢ 대부분의 augmentation 방식에서 좋은 성능을 보임

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 ±0.2	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

Ablation study on training methods on ImageNet

Experiments

❖ Fine-tuning at different resolution

- Training DeiT at resolution 224 x 224 and fine-tuning at resolution 384 x 384
- Fine-tuning 시에 positional embedding을 interpolation 수행
  - ➢ Bilinear interpolation은 L2-norm을 감소시켜서 오히려 accuracy가 떨어짐
  - ➢ 따라서, DeiT는 Bicubic interpolation을 사용함

Image Resolution

image size	throughput (image/s)	Imagenet [42] acc. top-1	Real [5] acc. top-1	V2 [41] acc. top-1
160^2	609.31	79.9	84.8	67.6
224^2	291.05	81.8	86.7	71.5
320^2	134.13	82.7	87.2	71.9
384^2	85.87	83.1	87.7	72.4

Experiments

❖ Training time

- DeiT-B 모델을 300 epoch만큼 training 할 때, 37시간 with 2 nodes 또는 53시간 with single nodes 소요
- DeiT-S와 DeiT-Ti는 4개의 GPU로 3일 이하 소요
  - ➢ 384 x384의 고해상도 이미지로 25 epoch만큼 fine-tuning하는 경우 (FixDeiT-B model), 20시간 with single node (8 GPU) 소요
- DeiT는 반복적인 augmentation을 3번 반복하여 사용하기 때문에 single epoch동안 전체 이미지의 1/3 만큼을 사용

https://www.youtube.com/watch?v=DjEvzeiWBTo

Conclusion

❖ Conclusion

- DeiT models can achieve competitive results against the state of the art on ImageNet with no external data. They are learned on a single node with 4 GPUs in three days. DeiT-S and DeiT-Ti have fewer parameters and can be seen as the counterpart of ResNet-50 and ResNet-18.
- With distillation procedure based on a distillation token, image transformers learn more from a convnet than from another transformer with comparable performance
- Transformer-specific strategy which uses new distillation procedure introduced in this paper outperforms vanilla distillation by a significant margin.
- DeiT models are competitive when transferred to different downstream tasks such as fine-grained classification, on several popular public benchmarks: CIFAR-10, CIFAR-100, Oxford-102 flowers, Stanford Cars and iNaturalist-18/19

Reference

1. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021, July). Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning (pp. 10347-10357). PMLR.


Thank You