Supplementary material

Training data-efficient image transformers & distillation through attention

This supplemental material provides complementary tables referred in the main document, in particular a more detailed description of the datasets that we used for transfer learning in Table 8. Table 9 provides the hyper-parameters used for Imagenet1k training. Table 10 compares different complexity measures of DeiT against state-of-the-art convnets and transformers. We also provide the PyTorch **code** associated with our paper in the supplemental material.

Table 8. Datasets used for our different tasks.

Dataset	Train size	Test size	#classes
ImageNet (Russakovsky et al., 2015)	1,281,167	50,000	1000
iNaturalist 2018 (Horn et al., 2018)	437,513	24,426	8,142
iNaturalist 2019 (Horn et al., 2019)	265,240	3,003	1,010
Flowers-102 (Nilsback & Zisserman, 2008)	2,040	6,149	102
Stanford Cars (Krause et al., 2013)	8,144	8,041	196
CIFAR-100 (Krizhevsky, 2009)	50,000	10,000	100
CIFAR-10 (Krizhevsky, 2009)	50,000	10,000	10

Table 9. Ingredients and hyper-parameters for Vit-B (Dosovitskiy et al., 2020) and our method.

Methods	ViT-B	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

Table 10. Measures of efficiency of our DeiT models with transformers and convnets architecture.

Model	Top-1	#params	FLOPs	im/s GPU	im/s GPU	im/s GPU	im/s GPU	im/s CPU	im/s CPU	GPU mem. used
	acc.	$\times 10^6$	$\times 10^9$	(BS=1,fp16)	(BS=1,fp32)	(BS=32,fp16)	(BS=32,fp32)	(BS=1)	(BS=32)	(BS=32,fp32)
EfficientNet B7	84.3	66	37.0	20.5	26.1	90.8	54.3	0.6	0.5	6207.0 MB
ViT-B	77.9	86	55.4	76.8	71.6	192.8	89.3	2.2	1.9	1686.2 MB
DeiT-Ti⚗	76.6	6	1.2	78.5	99.0	2363.2	2386.4	37.3	84.9	97.6 MB
DeiT-S⚗	82.6	22	4.6	76.5	95.5	1693.0	942.5	15.1	26.9	217.5 MB
DeiT-B⚗	84.2	87	17.5	79.5	95.0	745.5	303.9	5.5	7.1	579.2 MB
DeiT-B⚗ ↑384	85.2	87	55.4	76.6	71.5	192.7	89.3	2.1	1.9	1693.7 MB