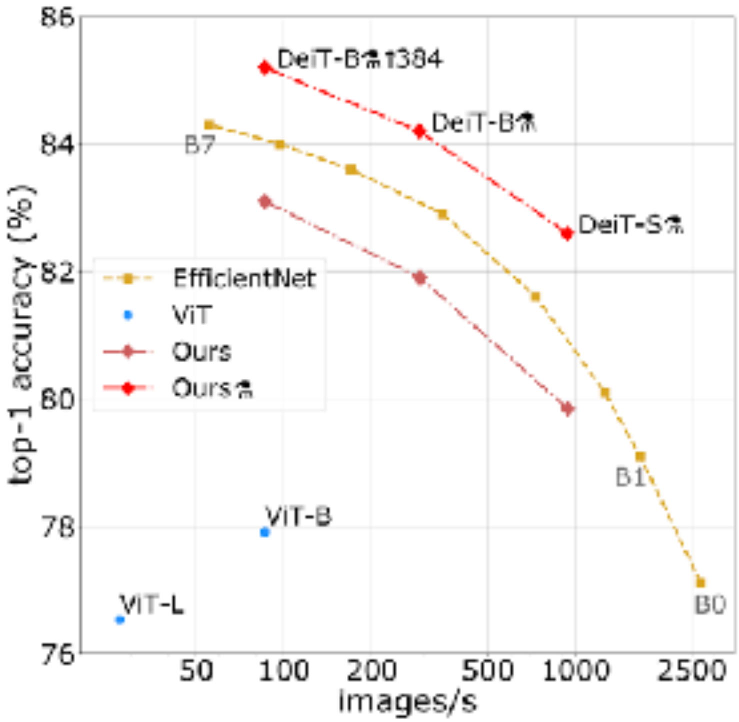
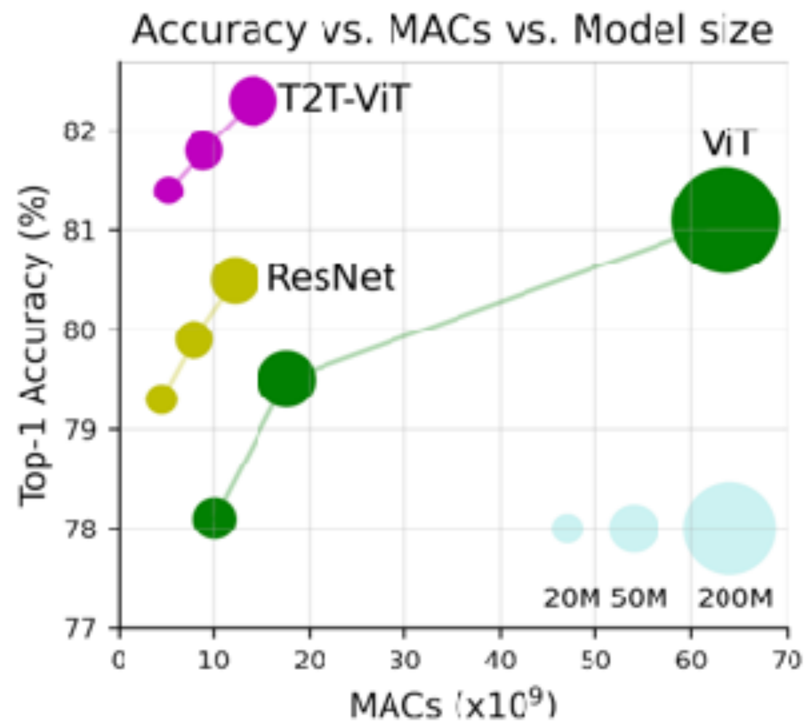
Paper Battle #1



vs.



DeiT [ICML'21]

T2T-ViT [ICCV'21]

Arguments for DeiT

Research Impact

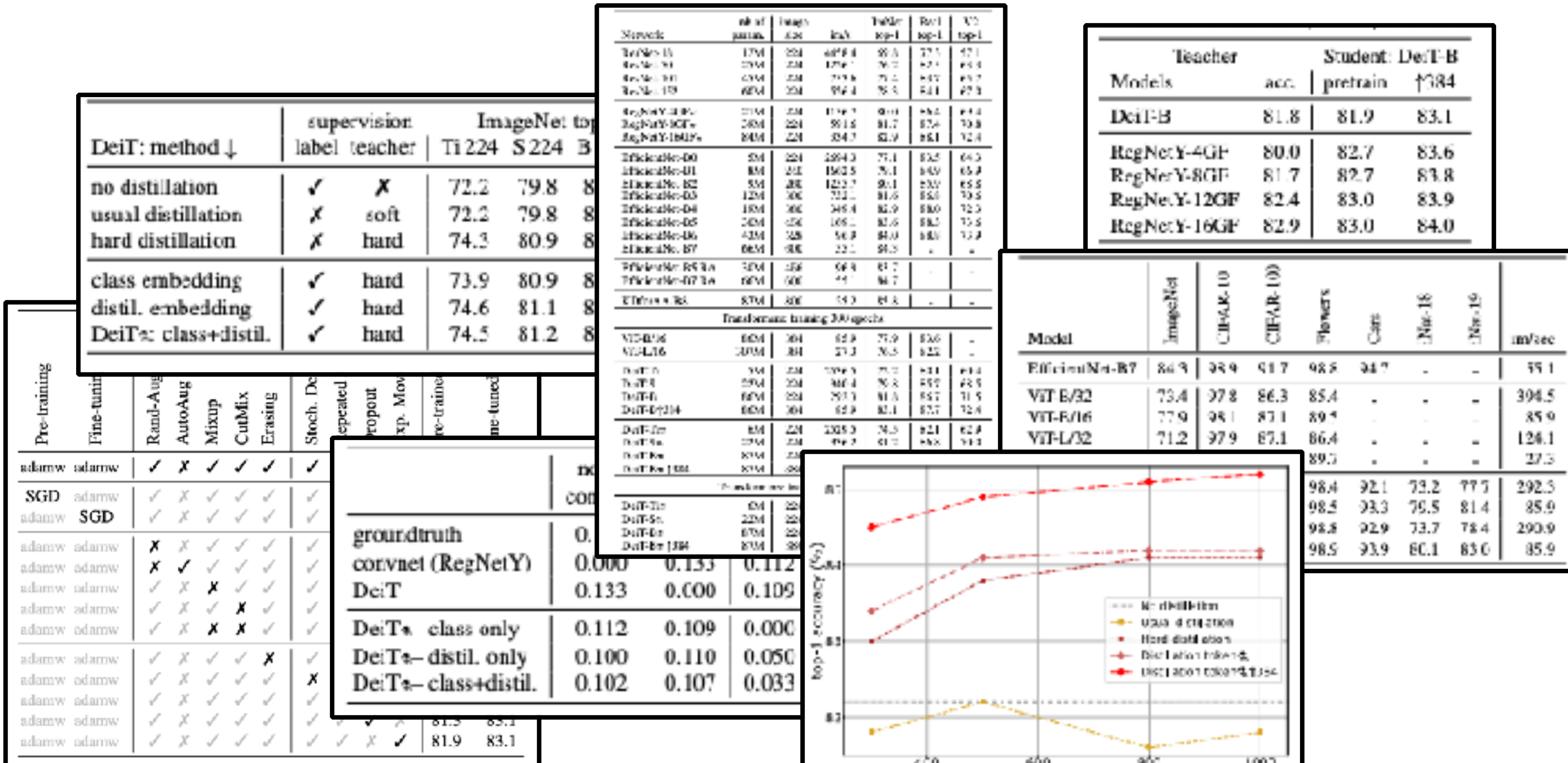
- Arguably, the DeiT paper had a larger impact on the visual recognition community than the T2T-ViT paper.

Training data-efficient image transformers & distillation through attention H Touvron, M Cord, M Douze, F Massa, A Sablayrolles, H Jégou International conference on machine learning, 10347-10357	4861	2021
Tokens-to-token vit: Training vision transformers from scratch on imagenet L Yuan, Y Chen, T Wang, W Yu, Y Shi, ZH Jiang, FEH Tay, J Feng, S Yan Proceedings of the IEEE/CVF international conference on computer vision, 558-567	1615	2021

deit	Public	Watch 51	Fork 535	Star 3.7k
T2T-ViT	Public	Watch 18	Fork 170	Star 1.1k

Empirical Insights

- The paper presents many thorough empirical studies, valuable for advancing the state-of-the-art in vision transformers.



DeiT: method ↓	supervision label	supervision teacher	ImageNet top Ti 224	S 224
no distillation	✓	✗	72.2	79.8
usual distillation	✗	soft	72.2	79.8
hard distillation	✗	hard	74.3	80.9
class embedding	✓	hard	73.9	80.9
distil. embedding	✓	hard	74.6	81.1
DeiT⚗: class+distil.	✓	hard	74.5	81.2

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	83.0	83.9
RegNetY-16GF	82.9	83.0	84.0

groundtruth	0.		
convnet (RegNetY)	0.000	0.133	0.112
DeiT	0.133	0.000	0.109
DeiT⚗ class only	0.112	0.109	0.000
DeiT⚗– distil. only	0.100	0.110	0.050
DeiT⚗– class+distil.	0.102	0.107	0.033

Model	ImageNet	CIFAR-10	CIFAR-100	Flowers	Cars	iNat-18	iNat-19	im/sec
ViT-B/32	73.4	97.8	86.3	85.4	-	-	-	394.5
ViT-L/32	71.2	97.9	87.1	86.4	-	-	-	124.1

Very Strong Results

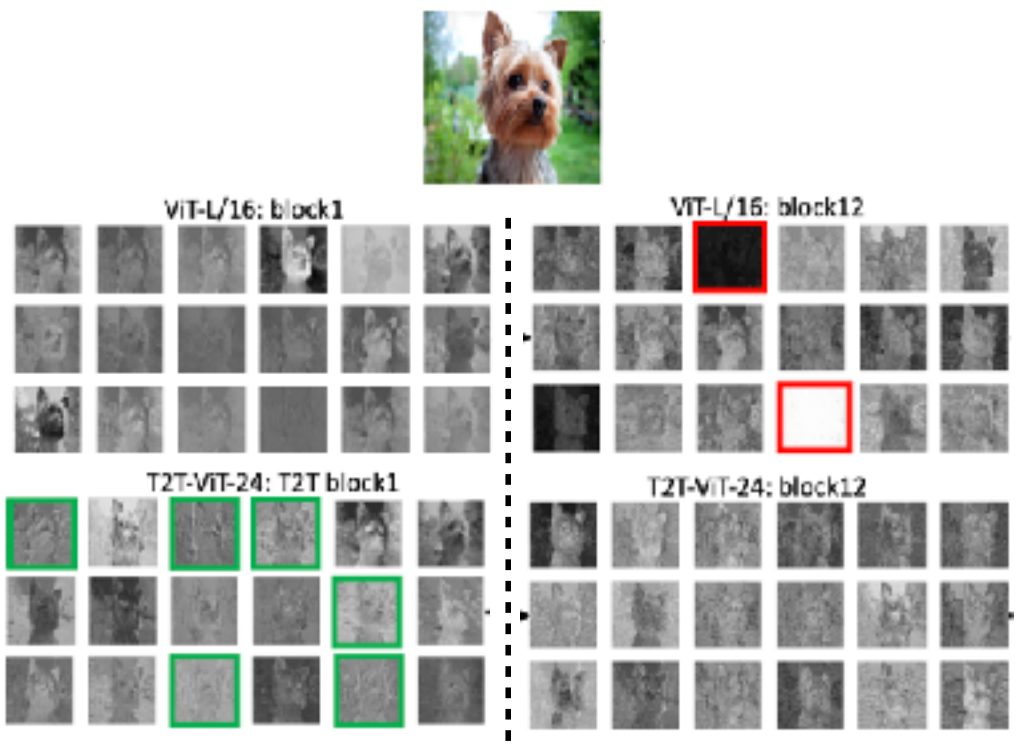
Network	nb of param.	image size	im/s	ImNet top-1	Real top-1	V2 top-1
ResNet-18	12M	224	4458.4	69.8	77.3	57.1
ResNet-50	25M	224	1226.1	76.2	82.5	63.3
ResNet-101	45M	224	753.6	77.4	83.7	65.7
ResNet-152	60M	224	526.4	78.3	84.1	67.0
RegNetY-4GF⋆	21M	224	1156.7	80.0	86.4	69.4
RegNetY-8GF⋆	39M	224	591.6	81.7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

The best DeiT-B model outperforms the best ViT-B model (even if the ViT is pretrained on the massive JFT dataset).

Arguments for T2T-ViT

A More Elegant Solution

- T2T systematically identifies two major problems of ViTs and proposes an elegant architectural solution to fix them.



a) Elegant solution of T2T-ViT

Pre-training	Fine-tuning	Rand-Augment	AutoAug	Mixup	CutMix	Erasing	Stoch. Depth	Repeated Aug.	Dropout	Exp. Moving Avg.	pre-trained 224	fine-tuned 384
adamw	adamw	✓	✗	✓	✓	✓	✓	✓	✗	✗	81.8 ±[illegible]	83.1 ±[illegible]
SGD	adamw	✓	✗	✓	✓	✓	✓	✓	✗	✗	74.5	77.3
adamw	SGD	✓	✗	✓	✓	✓	✓	✓	✗	✗	81.8	83.1
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
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

b) Brute force solution of DeiT

Better Results

- Compared to DeiT, T2T achieves higher accuracy without large CNN models as teachers to enhance the ViT.

Models	Top1-Acc (%)	Params (M)	MACs (G)
ViT-S/16 [12]	78.1	48.6	10.1
DeiT-small [36]	79.9	22.1	4.6
DeiT-small-Distilled [36]	81.2	22.1	4.7
T2T-ViT-14	**81.5**	21.5	4.8
T2T-ViT-14↑384	**83.3**	21.5	17.1
ViT-B/16 [12]	79.8	86.4	17.6
ViT-L/16 [12]	81.1	304.3	63.6
T2T-ViT-24	**82.3**	**64.1**	**13.8**

Impressive Accuracy vs Cost Tradeoff

- Compared to ResNets or ViTs, T2T achieves much better results for the same or even lower computational complexity.