Supplementary Material

A. Full results

We report ImageNet-1k top-1 accuracy and various cost indicators for every model configuration that appears in the figures of the main text (see Table A1, Table A2, Table A3). Throughput is measured on a single GeForce RTX 3090 GPU in mixed precision.

B. More implementation details

Hyperparameters. We train all of our models using the timm library [6] with the following hyperparameters: learning rate warmup for 5 epochs, learning rate cooldown for 10 epochs, cosine learning rate scheduler [3], weight decay 0.025, DropPath [2] rate 0.1, AdamW [4] optimizer with epsilon 1e-8, AutoAugment [1] image augmentations with configuration `rand-m9-mstd0.5-inc1`, mixup [8] alpha 0.8, cutmix [7] alpha 1.0, label smoothing 0.1. Unless otherwise specified, we use base learning rate 5e-5.

We fine-tune ViT-Small models for 130 epochs with batch size 1024, ViT-Base models for 60 epochs with batch size 400, and ViT-Large models for 20 epochs with batch size 192. For evaluation, we use exponential moving average (EMA) [5] with decay 0.99996. We use the default values in timm for all other hyperparameters.

ViT-Small

Method	#Patches	GMACs	Throughput	Runtime	ImageNet-1k
			ims/sec	μ-secs/im	Top-1 Acc.
Vanilla ViT	64	1.44	6489	154	74.55
	81	1.83	5208	192	76.36
	100	2.28	4212	237	77.55
	121	2.78	3460	289	78.26
	169	3.94	2315	432	79.84
	196	4.62	1975	506	80.28
Quadformer Feature-based scorer	64	1.54	3611	277	76.53
	79	1.88	3204	312	77.53
	100	2.37	2766	362	78.64
	121	2.87	2419	413	79.35
	169	4.04	1792	558	80.43
	196	4.71	1576	635	80.84
Quadformer Pixel-blur scorer	64	1.45	5150	194	74.97
	79	1.79	4362	229	76.27
	100	2.28	3590	279	77.47
	121	2.78	3022	331	78.58
	169	3.95	2104	475	80.01
	196	4.62	1813	552	80.4

Table A1. Full results - ViT Small.

ViT-Base

Method	#Patches	GMACs	Throughput	Runtime	ImageNet-1k
			ims/sec	μ-secs/im	Top-1 Acc.
Vanilla ViT	64	5.6	2676	374	80.78
	81	7.2	2155	464	81.73
	100	8.8	1739	575	82.31
	121	10.7	1429	700	82.71
	169	15.1	966	1035	83.74
	196	17.6	823	1215	84.07
Quadformer Feature-based scorer	64	5.7	2019	495	81.52
	79	7.1	1732	577	82.34
	100	8.9	1435	697	83.05
	121	10.8	1218	821	83.50
	169	15.2	864	1157	84.23
	196	17.7	750	1333	84.38
Quadformer Pixel-blur scorer	64	5.7	2424	413	80.78
	79	7.0	2021	495	81.68
	100	8.8	1630	613	82.57
	121	10.7	1354	739	83.06
	169	15.1	931	1074	83.87
	196	17.6	800	1250	84.23
Quadformer Oracle scorer	64	—	—	—	84.76
	79	—	—	—	85.19
	100	—	—	—	85.40
	121	—	—	—	85.67
	169	—	—	—	85.40
	196	—	—	—	85.25

Table A2. Full results - ViT Base.

ViT-Large

Method	#Patches	GMACs	Throughput	Runtime	ImageNet-1k
			ims/sec	μ-secs/im	Top-1 Acc.
Vanilla ViT	64	19.9	900	1111	82.00
	81	25.2	720	1389	83.02
	100	31.1	580	1724	83.86
	121	37.7	478	2092	84.46
	169	53.0	323	3096	85.42
	196	61.7	277	3610	85.74
Quadformer Feature-based scorer	64	20.1	777	1287	82.88
	79	24.7	649	1541	83.67
	100	31.3	527	1898	84.41
	121	37.9	440	2273	85.03
	169	53.1	306	3268	85.65
	196	61.8	265	3774	85.79
Quadformer Pixel-blur scorer	64	19.9	869	1151	81.66
	79	24.6	712	1404	82.69
	100	31.1	568	1761	83.61
	121	37.7	470	2128	84.3
	169	53.0	320	3125	85.22
	196	61.7	275	3636	85.56
Quadformer Oracle scorer	64	—	—	—	85.89
	79	—	—	—	86.33
	100	—	—	—	86.5
	121	—	—	—	86.7
	169	—	—	—	86.52
	196	—	—	—	86.54

Table A3. Full results - ViT-Large.

References

[1] Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. *ArXiv*, abs/1805.09501, 2018. 1

[2] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. In *European Conference on Computer Vision*, 2016. 1

[3] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv: Learning*, 2016. 1

[4] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 1

[5] Boris Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *Siam Journal on Control and Optimization*, 30:838–855, 1992. 1

[6] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 1

[7] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Young Joon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6022–6031, 2019. 1

[8] Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ArXiv*, abs/1710.09412, 2017. 1