

A Detailed Hyperparameters

We train all models with distributed data parallel (DDP) across devices and expert-parallelism within layers using Tutel (Hwang et al., 2023b) for efficient MoE implementation.

Hyperparameter Dataset	ViT Touvron et al. (2022)		ConvNeXt Liu et al. (2022)	
	IN1k	IN22k	IN1k	IN21k
Batch size	2048	2048	2048	2048
Optimizer	LAMB	LAMB	AdamW	AdamW
LR	$3 \cdot 10^{-3}$	$2.5 \cdot 10^{-3}$	$2.5 \cdot 10^{-3}$	$2.5 \cdot 10^{-3}$
LR decay	cosine	cosine	cosine	cosine
Weight decay	0.03	0.02	0.05	0.05
Expert Weight d.	0.06	0.04	0.05	0.05
Warmup epochs	5	5	20	5
Load balancing	0.01	0.01	0.01	0.01
Label smoothing ϵ	0.0	0.1	0.1	0.1
Stoch.Depth	✓	✓	✓	✓
RepeatedAug	✓	✗	✗	✗
H.flip	✓	✓	✓	✓
RRC	✗	✓	✗	✗
Rand Augment	✗	✗	✓	✓
3Augment	✓	✓	✗	✗
Color Jitter	0.3	0.3	✗	✗
Mixup alpha	0.8	0.0	0.8	0.0
Cutmix alpha	1.0	1.0	1.0	1.0
Erasing prob.	✗	✗	0.25	✗
Test crop ratio	1.0	1.0	0.875	0.875
Loss	BCE	CE	CE	CE

Table 10: Summary of our training procedures with ImageNet-1k (IN1k) and ImageNet-21k (IN21k).

Architecture	ImageNet-1k	ImageNet-21k
ViT-(S/B)	0.1 / 0.2	0.0 / 0.1
ConvNext-(T/S/B)	0.1 / 0.3 / 0.6	0.0 / 0.0 / 0.1
ConvNext-(T/B) <i>iso.</i>	0.1 / 0.5	-

B MoE vs Dense Table

Table 11: Experiments with pretraining on ImageNet-21k, JFT-300M or JFT-3B: performance with/without MoE. In the case of ConvNeXt-T, we get 0.6 improvement, while increasing the number of parameters, but almost not the number of per-sample parameters. For ConvNeXt-S, there is a 0.3 improvement. There is an improvement for SwinV2 models, but these were starting lower. Overall, for strong/big models, there is no clear improvement. Results without citations correspond to our work. V-MoE results marked with a star (*) are results taken from their GitHub (tinyurl.com/yb8fze5u). See Fig. 4 for an overview. For every model size examined, we present the accuracy metrics obtained after similar numbers of training iterations. Accuracy (Acc.) refers to Top-1 accuracy.

Architecture	Pre-train dataset	#Params ($\times 10^6$)	Per samples #Params _{act}	IN-1K Acc.
ViT-s/32(Riquelme et al., 2021)	JFT-300M	36.5	36.5	73.73
ViT-s/32-32, Last 2(Riquelme et al., 2021)	JFT-300M	166.7	≈ 55	77.10
ViT-s/32-32, Every 2(Riquelme et al., 2021)	JFT-300M	296.9	≈ 70	77.10
ViT-S/16 (Touvron et al., 2022)	ImageNet-21k	22.0	22.0	82.6
ViT-S/16-8 Every 2 Top 2	ImageNet-21k	71.7.6	33.1	83.0
ConvNeXt-T (Liu et al., 2022)	ImageNet-21k	28.6	28.6	82.9
ConvNeXt-T-8 Last 2 Top 1	ImageNet-21k	70.0	28.7	83.5
SwinV2-S (Hwang et al., 2023a)	ImageNet-21k	65.8	-	83.5
SwinV2-S-8 (Hwang et al., 2023a)	ImageNet-21k	173.3	65.8	84.5
SwinV2-S-16 (Hwang et al., 2023a)	ImageNet-21k	296.1	65.8	84.9
SwinV2-S-32 (Hwang et al., 2023a)	ImageNet-21k	296.1	65.8	84.7
ConvNeXt-S (Liu et al., 2022)	ImageNet-21k	50.3	-	84.6
ConvNeXt-S-8 Last 2 Top 1	ImageNet-21k	91.6	50.3	84.9
ViT-B/16 (Touvron et al., 2022)	ImageNet-21k	86.6	86.6	85.2
ViT-B/16, Every 2, Top 2	ImageNet-21k	284.9	129.9	85.2
ViT-B/16 \uparrow 384 (Touvron et al., 2022)	ImageNet-21k	86.6	86.6	86.7
ViT-B/16 (Zhai et al., 2022)	JFT-300M	86.6	86.6	84.9
ViT-B/16 (Riquelme et al., 2021)	JFT-300M	100.5	100.5	84.15
V-MoE-B/16-32, Every 2 (Riquelme et al., 2021)	JFT-300M	979.0	≈ 200	85.3
V-MoE-B/16-32, Last 2 (Riquelme et al., 2021)	JFT-300M	393.3	≈ 110	85.4
SwinV2-B (Hwang et al., 2023a)	ImageNet-21k	109.3	109.3	85.1
SwinV2-B-8 (Hwang et al., 2023a)	ImageNet-21k	300.3	109.3	85.3
SwinV2-B-16 (Hwang et al., 2023a)	ImageNet-21k	518.7	109.3	85.5
SwinV2-B-32 (Hwang et al., 2023a)	ImageNet-21k	955.3	109.3	85.5
ConvNeXt-B (Liu et al., 2022)	ImageNet-21k	88.6	88.6	85.8
ConvNeXt-B-8	ImageNet-21k	162.0	88.6	85.7
ViT-L/16 (Touvron et al., 2022)	ImageNet-21k	304.4	304.4	87.0
ViT-L/16 (Zhai et al., 2022)	JFT-300M	323.1	323.1	87.7
ViT-L/16 (Riquelme et al., 2021)	JFT-300M	323.1	323.1	87.1
V-MoE-L/16-32, Every 2 (Riquelme et al., 2021)	JFT-300M	3446.0	≈ 600	87.4
V-MoE-L/16-32, Last 2 (Riquelme et al., 2021)	JFT-300M	843.6	≈ 380	87.5
SoViT/14(Alabdulmohsin et al., 2023)	JFT-3B	400	400	90.3
ViT-g/14(Zhai et al., 2022)	JFT-3B	1B	1B	90.45
V-MoE/14 (Riquelme et al., 2021)	JFT-3B	15B	$\approx 1B$	90.35

C Ade20k

Table 12: ImageNet-1K trained models

Backbone	Input crop. ($\times 10^6$)	#Params #Params	per sample	Single scale mIoU
ImageNet-1K trained models				
ConvNeXt-T Liu et al. (2022)	512 ²	28.6	-	46.0
ConvNeXt-T-4 Last 2 Top 1	512 ²	34.5	25.6	46.0
ConvNeXt-S Liu et al. (2022)	512 ²	50	-	48.7
ConvNeXt-S-4 Last 2 Top 1	512 ²	56.1	47.3	48.38
ConvNeXt-B Liu et al. (2022)	512 ²	88.6	-	49.1
ConvNeXt-B-4 Last 2 Top 1	512 ²	99.1	83.4	48.8
ViT-S Touvron et al. (2022)	512 ²	88.6	-	45.8
ViT-S-8 Every 2 Top 2	512 ²	99.1	83.4	45.4
ViT-B Touvron et al. (2022)	512 ²	88.6	-	49.0
ViT-B-8 Every 2 Top 2	512 ²	99.1	83.4	48.3
ImageNet-22K trained models				
ConvNeXt-T Liu et al. (2022)	640 ²	28.6	-	48.6
ConvNeXt-T-4 Last 2 Top 1	640 ²	34.5	25.6	48.9
ConvNeXt-S Liu et al. (2022)	640 ²	50	-	51.0
ConvNeXt-S-4 Last 2 Top 1	640 ²	56.1	47.3	50.8
ConvNeXt-B Liu et al. (2022)	640 ²	88.6	-	52.6
ConvNeXt-B-4 Last 2 Top 1	640 ²	99.1	83.4	52.0
ViT-S Touvron et al. (2022)	640 ²	88.6	-	48.4
ViT-S-8 Every 2 Top 2	640 ²	99.1	83.4	48.7
ViT-B Touvron et al. (2022)	640 ²	88.6	-	52.8
ViT-B-8 Every 2 Top 2	640 ²	99.1	83.4	52.4