APPENDIX

Robustness Evaluation

We check the generalization of our NEXcepTion models on three computer vision robustness benchmarks: ImageNet-A (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al., 2021a) and ImageNet-Sketch (Wang et al., 2019). We use the models trained on ImageNet-1K, without any additional data or training. A comparison of the accuracies on the robustness benchmarks is available in Table 3. We distinguish how all the NEXcepTion models outperform the original Xception in all the robustness benchmarks and obtain results similar to the current state-of-the-art ConvNets and Transformers. Indeed, the variant NEXcepTion-S surpasses all the other compared networks on ImageNet-Sketch.

Architectures

In this section we present the details of the NEXcepTion architectures. In Table 4 we compare ResNet-50 and ConvNeXt-T with our model NEXcepTion-TP, which is very similar to ConvNeXt-T. In NEXcepTion-TP we replace the ConvNeXt block with a NEXcepTion block, and add one more block to match the same number of FLOPs. In Table 5 we depict the differences between the original Xception architecture and the architecture of NEXcepTion-T. In Table 6 the differences when compared to NEXcepTion-S are shown. This variant is similar to NEXcepTion-T with more channels to match the number of FLOPs of the original Xception.

Table 1: Classification accuracy on ImageNet-1K. χ for Convolutional and τ for Transformer networks. We present the results sorted by their FLOPs value. The values are obtained from the cited publications, except for the throughput's, calculated using the *timm* library on a RTX 2080 Ti during 30 repetitions.

[χ/τ] model	input img.	#params	FLOPs	throughput (images / s)	IN-1K top-1 acc.
χ EffNet-B1 (Tan and Le, 2019)	240^2	7.8M	0.7G	1618 ± 24	79.1
χ EffNet-B2 (Tan and Le, 2019)	260^2	9.2M	1.0G	1308 ± 9	80.1
χ EffNet-B3 (Tan and Le, 2019)	300^2	12M	1.8G	867 ± 6	81.6
τ NAT-Mini (Hassani et al., 2022)	224^2	20M	2.7G	713 ± 3	81.8
χ ResNet-50 (Wightman et al., 2021)	224^2	25.6M	4.1G	1633 ± 26	80.4
χ EffNet-B4 (Tan and Le, 2019)	380^2	19M	4.2G	937 ± 6	82.9
χ ConvNeXt-S (iso.) (Liu et al., 2022)	224^2	22M	4.3G	1677 ± 77	79.7
τ NAT-T (Hassani et al., 2022)	224^2	28M	4.3G	527 ± 1	83.2
τ Swin-T (Liu et al., 2021b)	224^2	28M	4.5G	867 ± 7	81.3
χ NEXcepTion-TP	224^2	26.6M	4.5G	1428 ± 9	81.8
χ ConvNeXt-T (Liu et al., 2022)	224^2	29M	4.5G	1125 ± 5	82.1
τ ViT-S (Dosovitskiy et al., 2021; Liu et al., 2022)	224^2	22M	4.6G	1330 ± 8	79.8
τ DeiT-S (Touvron et al., 2021)	224^2	22M	4.6G	1332 ± 13	79.8
χ NEXcepTion-T	224^2	24.5M	4.7G	965 ± 6	81.5
τ NAT-S (Hassani et al., 2022)	224^2	51M	7.8G	359 ± 1	83.7
χ ResNet-101 (Wightman et al., 2021)	224^2	44.5M	7.9G	1157 ± 6	81.5
χ Xception (Chollet, 2017)	299^2	23.6M	8.4G	756 ± 5	79.0
χ NEXcepTion-S	224^2	43.4M	8.5G	772 ± 3	82.0
τ Swin-S (Liu et al., 2021b)	224^2	50M	8.7G	472 ± 1	83.0
χ ConvNeXt-S (Liu et al., 2022)	224^2	50M	8.7G	753 ± 3	83.1
χ EffNetV2-S (Tan and Le, 2021)	384^2	22M	8.8G	543 ± 6	83.9
χ EffNet-B5 (Tan and Le, 2019)	456^2	30M	9.9G	476 ± 2	83.6

Table 2: Training details of the proposed networks.

Training Configuration	NEXcepTion-S	NEXcepTion-T	NEXcepTion-TP
Input Size	224	224	224
Optimizer	LAMB	LAMB	LAMB
Learning Rate	$1.4e^{-3}$	$2e^{-3}$	$2e^{-3}$
Weight Decay	0.02	0.02	0.02
Batch Size	128	256	256
Training Epochs	300	300	300
Learning Rate Schedule	Cosine Decay	Cosine Decay	Cosine Decay
Warmup Epochs	5	5	5
RandAugment	7 / 0.5	7 / 0.5	7 / 0.5
Mixup	0.1	0.1	0.1
Cutmix	1.0	1.0	1.0
Random Erasing	0.0	0.0	0.0
Label Smoothing	0	0	0
Stochastic Depth	0.05	0.05	0.05
Minimum Learning Rate	$1.0e^{-06}$	$1.0e^{-06}$	$1.0e^{-06}$
Error Function	BCE	BCE	BCE
Test crop ratio	0.95	0.95	0.95

Table 3: Robustness evaluation of NEXcepTion compared to other state-of-the-art models.

Model	FLOPs / Params	ImageNet-A	ImageNet-R	ImageNet-Sketch
NEXcepTion-T	4.7 / 24.5	17.65	45.99	34.73
NEXcepTion-TP	4.5 / 26.6	22.75	47.37	34.70
NEXcepTion-S	8.5 / 43.4	21.36	47.82	36.60
Xception	8.4 / 43.4	9.83	40.79	29.88
Swin-T	4.5 / 28.3	21.60	41.30	29.10
ConvNeXt-T	4.5 / 28.6	24.20	47.20	33.80
RVT-S	4.7 / 23.3	25.70	47.70	34.70

Table 4: Detailed architecture specifications for ResNet-50, ConvNeXt-T and NEXcepTion-TP.

	output size	ResNet-50	ConvNeXt-T	NeXcepTion-TP
stem	56×56	7×7, 64, stride 2 3×3 max pool, stride 2	4×4, 96, stride 4	4×4, 96, stride 4
res2	56×56	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} d7\times7, 96 \\ 1\times1, 384 \\ 1\times1, 96 \end{bmatrix} \times 3$	$\begin{bmatrix} d5\times5, 288 \\ d5\times5, 96 \\ d5\times5, 96 \\ \text{SE Module} \end{bmatrix} \times 3$
res3	28×28	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} d7\times7, 192 \\ 1\times1, 768 \\ 1\times1, 192 \end{bmatrix} \times 3$	$\begin{bmatrix} d5\times5, 576 \\ d5\times5, 192 \\ d5\times5, 192 \\ \text{SE Module} \end{bmatrix} \times 4$
res4	14×14	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} d7\times7, 384 \\ 1\times1, 1536 \\ 1\times1, 384 \end{bmatrix} \times 9$	$\begin{bmatrix} d5\times5, 1152 \\ d5\times5, 384 \\ d5\times5, 384 \\ \text{SE Module} \end{bmatrix} \times 9$
res5	7×7	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} d7\times7, 768 \\ 1\times1, 3072 \\ 1\times1, 768 \end{bmatrix} \times 3$	$\begin{bmatrix} d5\times5, 2304 \\ d5\times5, 768 \\ d5\times5, 768 \\ \text{SE Module} \end{bmatrix} \times 3$
FLOPs		4.1×10^9	4.5×10^9	4.5×10^9
# params.		25.6×10^6	28.6×10^6	56.6×10^6

Table 5: Architecture comparison between Xception and NEXcepTion-T. In the right side of each model the residual layers are described for each block. All the convolutions are depthwise separable convolutions except for the convolutions on the stem block and on the residual connections. The MaxPool operations have stride 2.

	Xception		NEXcepTion-T	
Input size	$299 \times 299 \times 3$		$224 \times 224 \times 3$	
Stem	3×3, 32, $s = 2$ 3×3, 64		2×2, 96, $s = 2$	
Entry flow	3×3, 128 3×3, 128 MaxPool 3×3	| [1×1, s=2] |	5×5, 128 5×5, 128 MaxBlurPool 3×3 SE Module	| [1×1, s=2] | |
	3×3, 256 3×3, 256 MaxPool 3×3	| [1×1, s=2] |	5×5, 256 5×5, 256 MaxBlurPool 3×3 SE Module	| [1×1, s=2] | |
	3×3, 728 3×3, 728 MaxPool 3×3	| [1×1, s=2] |	5×5, 512 5×5, 512 MaxBlurPool 3×3 SE Module	| [1×1, s=2] | |
Resulting feature maps	$19 \times 19 \times 728$		$14 \times 14 \times 512$	
Middle flow $\times$ 8	3×3, 728 3×3, 728 3×3, 728	| $\times 1$ |	5×5, 1536 5×5, 512 5×5, 512 SE Module	| $\times 1$ | |
Resulting feature maps	$19 \times 19 \times 728$		$14 \times 14 \times 512$	
Exit flow	3×3, 728 3×3, 1024 MaxPool 3×3	| [1×1, s=2] |	5×5, 512 5×5, 1024 MaxBlurPool 3×3 SE Module	| [1×1, s=2] | |
	3×3, 1536 3×3, 2048 Global Average Pooling		3×3, 1536 3×3, 2048 Global Average Pooling	
Output	Fully connected layers		Fully connected layers	

Table 6: Architecture comparison between Xception and NEXcepTion-S. In the right side of each model the residual layers are described for each block. All the convolutions are *depthwise separable convolutions* except for the convolutions on the stem block and on the residual connections. The MaxPool operations have stride 2.

	Xception		NEXcepTion-S	
Input size	$299 \times 299 \times 3$		$224 \times 224 \times 3$	
Stem	3×3, 32, $s = 2$ 3×3, 64		2×2, 96, $s = 2$	
Entry flow	3×3, 128 3×3, 128 MaxPool 3×3	| [1×1, s=2] |	5×5, 128 5×5, 128 MaxBlurPool 3×3 SE Module	| [1×1, s=2] | |
	3×3, 256 3×3, 256 MaxPool 3×3	| [1×1, s=2] |	5×5, 256 5×5, 256 MaxBlurPool 3×3 SE Module	| [1×1, s=2] | |
	3×3, 728 3×3, 728 MaxPool 3×3	| [1×1, s=2] |	5×5, 752 5×5, 752 MaxBlurPool 3×3 SE Module	| [1×1, s=2] | |
Resulting feature maps	$19 \times 19 \times 728$		$14 \times 14 \times 752$	
Middle flow $\times$ 8	3×3, 728 3×3, 728 3×3, 728	| $\times 1$ |	5×5, 2256 5×5, 752 5×5, 752 SE Module	| $\times 1$ | |
Resulting feature maps	$19 \times 19 \times 728$		$14 \times 14 \times 752$	
Exit flow	3×3, 728 3×3, 1024 MaxPool 3×3	| [1×1, s=2] |	5×5, 752 5×5, 1024 MaxBlurPool 3×3 SE Module	| [1×1, s=2] | |
	3×3, 1536 3×3, 2048 Global Average Pooling		3×3, 1536 3×3, 2048 Global Average Pooling	
Output	Fully connected layers		Fully connected layers	