Revisiting Adversarial Training for ImageNet: Architectures, Training and Generalization across Threat Models

Naman Deep Singh*, Francesco Croce* & Matthias Hein

Eberhard Karls Universität Tübingen

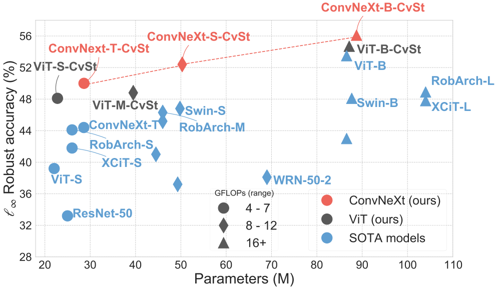
Tübingen AI Center
tuebingen.ai

imprs-is

Problem: adversarial training (AT) is well-studied on small-scale datasets (CIFAR) and ResNets, but **not explored in the modern setup** (ImageNet, new architectures)

Our contributions:

- We **revisit AT on ImageNet** and propose a training scheme effective across architectures.



- We achieve **SOTA robust classifiers** w.r.t. ℓ_∞. And show that using **ConvStem** instead of PatchStem boosts the generalization to the unseen ℓ_1 and ℓ_2 threat models.
- We uncover a **surprising** phenomena: increasing **test-time resolution** enhances robustness.

Proposed Adversarial Training setup

Aspects considered: **initialization, augmentations and training time/epochs**.

We do 2-step AT with APGD for the ℓ_∞-threat model at radius 4/255, and never train for ℓ_2 and ℓ_1 threat models.

Using strong pre-trained models as initialization

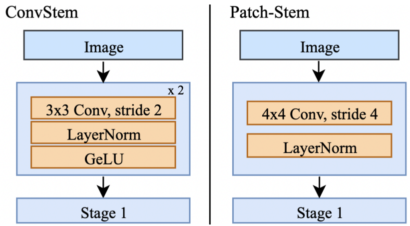
- yields more robust models
- allows us to use **heavy augmentations** which in particular with **longer training** yields robustness gains, contrary to suggestions from prior work **[1]**.

Architecture	Training Scheme	Adversarial Training w.r.t. ℓ_∞ clean	ℓ_∞	ℓ_2	ℓ_1
ConvNeXt-T	random init., basic augment.	64.0	37.1	28.9	8.8
	+ strong clean pre-training	69.4 +5.4	41.6 +4.5	42.4 +13.5	18.9 +10.1
	+ heavy augmentations	71.0 +1.6	46.5 +4.9	38.1 -4.3	14.9 -4.0
	50 $\rightarrow$ 300 epochs	72.4 +1.4	48.6 +2.1	38.0 -0.1	14.9 +0.0
ViT-S	random init., basic augment.	61.5	31.8	35.5	15.1
	+ strong clean pre-training	66.8 +5.3	39.1 +7.3	34.5 -1.0	12.6 -2.5
	+ heavy augmentations	65.2 -1.6	39.2 +0.1	37.3 +2.8	16.2 +3.6
	50 $\rightarrow$ 300 epochs	69.2 +4.0	44.0 +4.8	37.5 +0.2	15.1 -1.1

Both **ViTs and ConvNeXts have a PatchStem**, which has strong downsampling in a single layer.

Use ConvStem instead:

- Downsamplings spread across multiple layers.
- Longer training also possible without **overfitting**.
- **ConvStem** improves ℓ_∞ robustness and boosts generalization to unseen threat models.



Architecture	Adversarial training w.r.t ℓ_∞			
	clean	ℓ_∞	ℓ_2	ℓ_1
ConvNeXt-T	72.4	48.6	38.0	14.9
ConvNeXt-T + CvSt	72.7 +0.3	49.5+0.9	48.4 +10.4	24.5 +9.6
ViT-S	69.2	44.0	37.5	15.1
ViT-S + CvSt	72.5 +3.3	48.1 +4.1	50.4 +12.9	26.7 +11.6

What is the proposed scheme for adversarial training for high-resolution datasets like ImageNet?

Strong pre-trained models as initialization

+

Heavy augmentations

+

longer training

+

Using a ConvStem

Does it scale?

Yes, even to Large models

We use the previously found training scheme and train models of size **Small** (<30M params), and **Large** (>80M params) for 250/300 epochs.

	Architecture	Params (M)	FLOPs (G)	Source	Ep.	Adv. Steps	Adversarial Tr. wrt ℓ_∞ clean	ℓ_∞	ℓ_2	ℓ_1
Small	ResNet-50	25.0	4.1	[46]	100	3	65.88	33.18	18.88	3.82
	XCiT-S12	26.0	4.8	[14]	110	1	72.34	41.78	46.20	22.72
	ViT-S	22.1	4.6	ours	300	2	69.22	44.04	37.52	15.12
	RobArch-S	26.1	6.3	[42]	110	3	70.58	44.12	39.88	15.46
	Isotropic-CN-S	22.3	4.3	ours	300	2	69.04	44.22	36.64	14.88
	ConvNeXt-T	28.6	4.5	[14]	110	1	71.60	44.40	45.32	21.76
	ViT-S + ConvStem	22.8	5.0	ours	300	2	72.56	48.08	**50.40**	26.68
	ConvNeXt-T	28.6	4.5	ours	300	2	72.40	48.60	38.02	14.88
	ConvNeXt-T + ConvStem	28.6	4.6	ours	300	2	**72.72**	49.46	48.42	24.52
Large	Swin-B	87.7	15.5	[37]	90	3	74.76	48.10	44.42	18.04
	RobArch-L	104.0	25.7	[42]	100	3	73.46	48.92	39.48	14.74
	ViT-B	86.6	17.6	[44]	300	2	**76.62**	53.50	–	–
	ViT-B + ConvStem	87.1	17.9	ours	250	2	76.30	54.66	**56.30**	32.06
	ConvNeXt-B	88.6	15.4	[32]*	300	3	76.02	55.82	44.68	21.23
	ConvNeXt-B + ConvStem	88.8	16.0	ours	250	2	75.90	56.14	49.12	23.34

Our **ConvNeXt + Convstem yields the most robust models to ℓ_∞** whereas **ViT + CovnStem gives the best generalization** to unseen threat models.

The training scheme scales, without overfitting to yield SOTA robust classifiers.

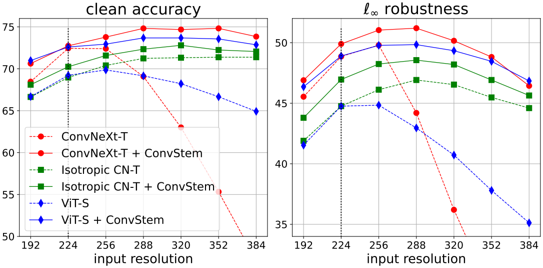
Our robust models are publically available.

Robust ImageNet models can be used for a lot of downstream tasks. We show that using these models as initialization allows us to train SOTA robust semantic segmentation models [4].

Increasing test-time resolution

Well known: Increasing resolution at test-time improves clean performance.
Is this true also for robust accuracy despite the threat model becoming more powerful?

- **Surprisingly**, increasing input-resolution at test-time also improves robustness, with diminishing returns for very high resolutions.
- **Remark:** We never train at the increased resolution.



- **ConvStem models are more stable** (their decay for higher resolutions is not as severe) compared to PatchStem models.
- Also seen for perturbation radius larger than training (6/255 and 8/255).

References

[1] Edoardo Debenedetti, Vikash Sehwag, and Prateek Mittal. **"A Light Recipe to Train Robust Vision Transformers".** In: *First IEEE Conference on Secure and Trustworthy Machine Learning*. 2023

[2] Chang Liu et al. **"A Comprehensive Study on Robustness of Image Classification Models: Benchmarking and Rethinking".** In: *arXiv preprint, arXiv:2302.14301* (2023)

[3] Sylvestre-Alvise Rebuffi, Francesco Croce, and Sven Gowal. **"Revisiting adapters with adversarial training".** In: *ICLR*. 2023

[4] Francesco Croce, Naman D Singh, and Matthias Hein. **"Robust Semantic Segmentation: Strong Adversarial Attacks and Fast Training of Robust Models".** In: *arXiv preprint, arXiv:2306.12941* (2023)