A *Light* Recipe To Train Robust Vision Transformers

Edoardo Debenedetti - ETH Zürich & EPFL
Vikash Sehwag - Princeton University
Prateek Mittal - Princeton University

edebenedetti@inf.ethz.ch
@edoardo_debe

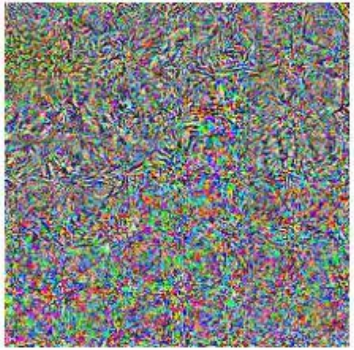
Raleigh, North Carolina, USA - 8/2/2023 - IEEE SaTML 2023

Paper

Code

Adversarial Examples

$+ \varepsilon \times$



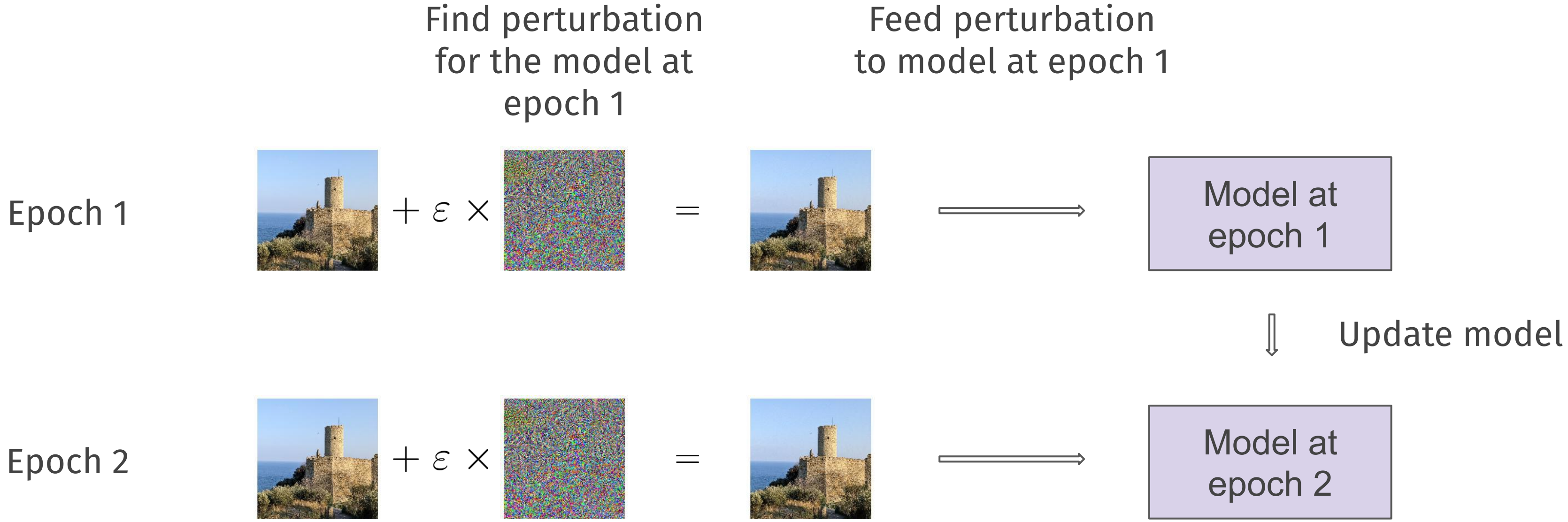
$=$

"Castle" $\hat{\delta}$ "Bee"

[1] Szegedy, Christian, et al. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).
[2] Biggio, Battista, et al. "Evasion attacks against machine learning at test time." Joint European conference on machine learning and knowledge discovery in databases. Springer, Berlin, Heidelberg, 2013.

A solution: adversarial training

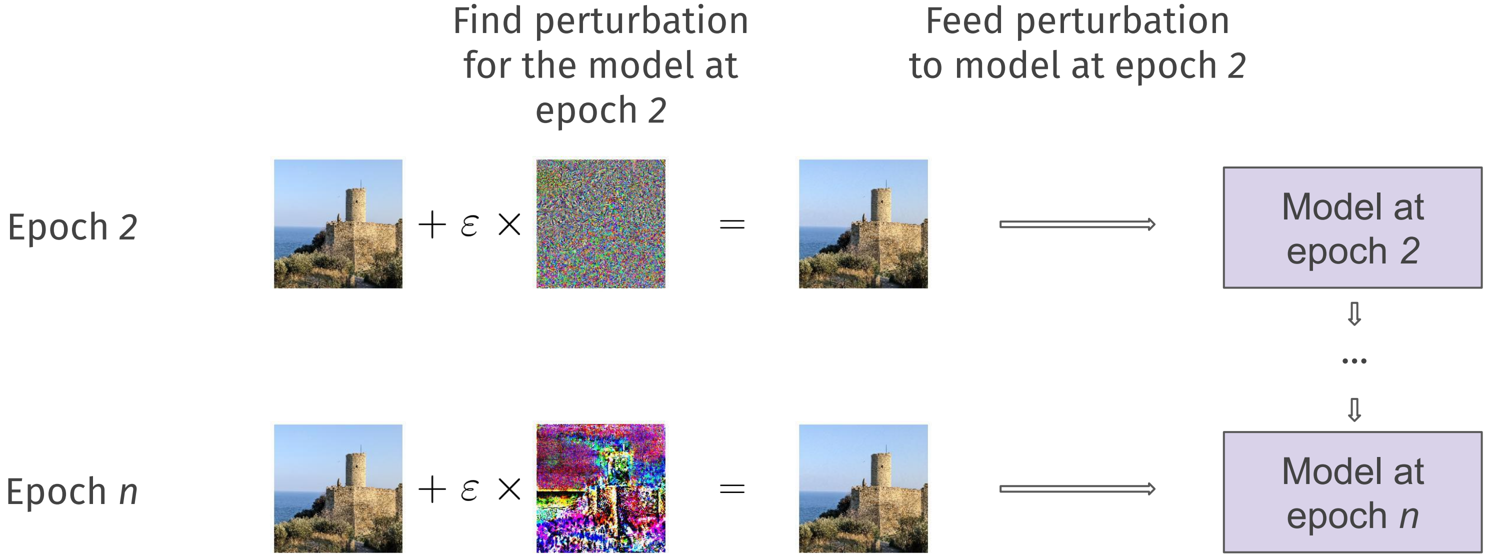
- Train on adversarial examples instead of clean data
- Theoretically principled and effective in practice



[3] Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." arXiv preprint arXiv:1706.06083 (2017).

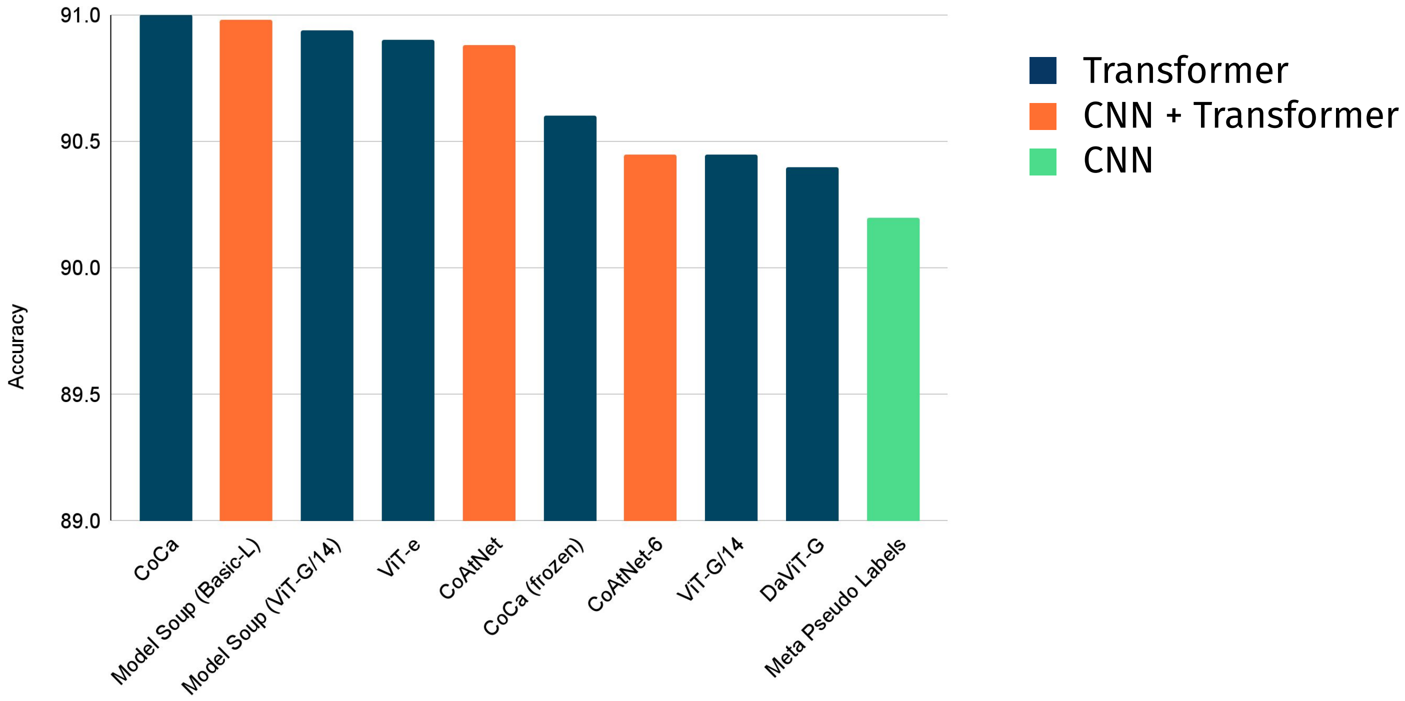
A solution: adversarial training

- Train on adversarial examples instead of clean data
- Theoretically principled and effective in practice



[3] Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." arXiv preprint arXiv:1706.06083 (2017).

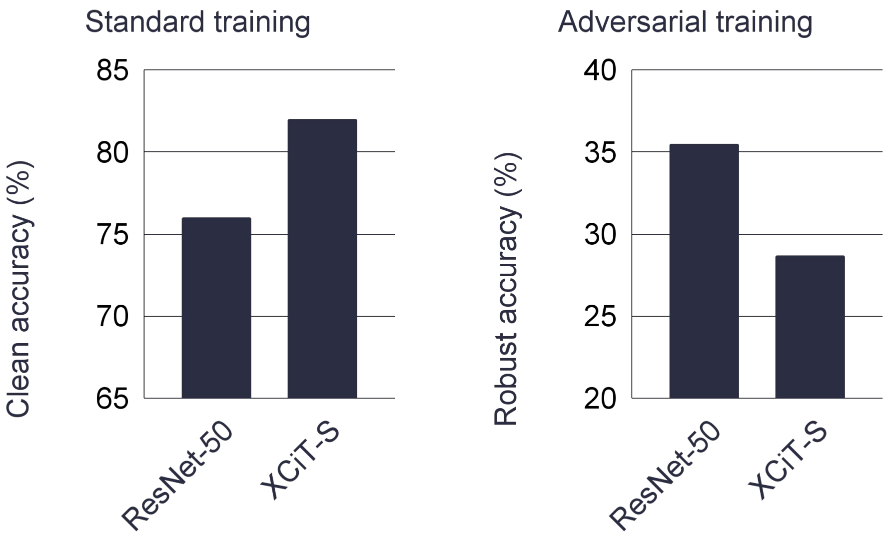
The Vision Transformers (ViTs) family is here!



Data in the plot from PapersWithCode: https://paperswithcode.com/sota/image-classification-on-imagenet, retrieved on 2022/2/6.

[4] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

Are ViTs good at adversarial training?



Despite being better than ResNet-50 in terms of clean accuracy when standardly trained, XCiT-S performs worse if trained with adversarial training.

[5] Ali, Alaaeldin, et al. "Xcit: Cross-covariance image transformers." Advances in neural information processing systems 34 (2021).
[6] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

Contributions of this work

- Vision Transformers can be competitive at adversarial training, but need a custom adversarial training recipe
- Our recipe generalizes to larger variants and different architectures
- One potential reason of why the recipe matters so much: it influences the inner part of adversarial training

Contributions of this work

- Vision Transformers can be competitive at adversarial training, but need a custom adversarial training recipe
- Our recipe generalizes to larger variants and different architectures
- One potential reason of why the recipe matters so much: it influences the inner part of adversarial training

Set-up

- Adversarially train on ImageNet using 1-step FGSM for L-∞ perturbations with ε = 4/255
- Start from the standard training set-up of DeiT
- Search for optimal parameters in terms of:
  - Architecture
  - Warming up attack strength
  - Data augmentation
  - Weight decay
- Evaluate using AutoAttack (but the intermediate ablations with the faster APGD-CE)

[7] Wong, Eric, Leslie Rice, and J. Zico Kolter. "Fast is better than free: Revisiting adversarial training." arXiv preprint arXiv:2001.03994 (2020).
[8] Touvron, Hugo, et al. "Training data-efficient image transformers & distillation through attention." International Conference on Machine Learning. PMLR, 2021.
[9] Croce, Francesco, and Matthias Hein. "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks." International conference on machine learning. PMLR, 2020.

Finding the optimal recipe: Bag-of-tricks

Feature	Accuracy	
	Clean	AutoAttack
XCiT-S12	71.68	28.70
+ ε warmup (10 epochs)	71.98 (+0.30)	29.36 (+0.66)
+ Tuned data augmentation	71.70 (−0.28)	38.78 $\boxed{(+9.42)}$
+ Tuned weight decay	**72.34** (+0.64)	**41.78** (+3.00)

Summary of the improvements given by each phase of the ablation. Overall, we improve the robust accuracy by **13.08%**, and the clean one by **0.66%** over the baseline.

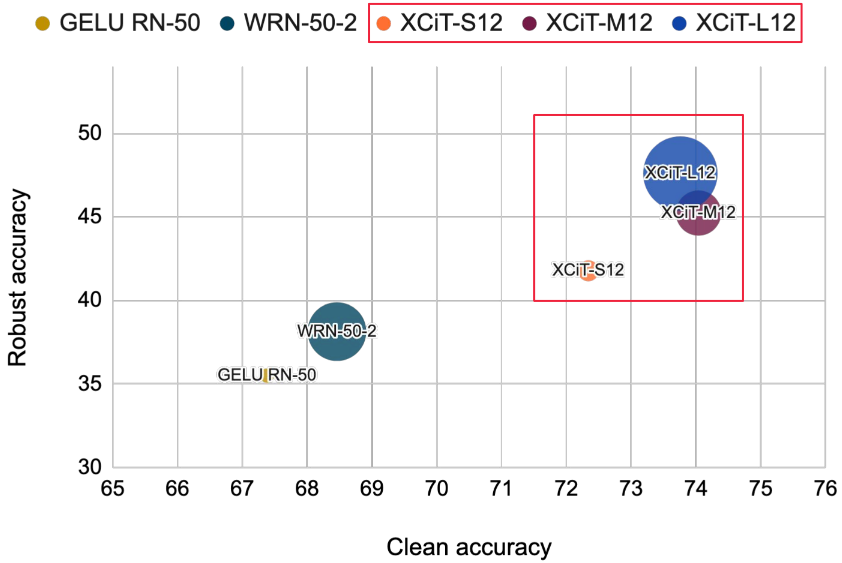
Contributions of this work

- Vision Transformers can be competitive at adversarial training, but need a custom adversarial training recipe
- Our recipe generalizes to larger variants and different architectures
- One potential reason of why the recipe matters so much: it influences the inner part of adversarial training

Contributions of this work

- Vision Transformers can be competitive at adversarial training, but need a custom adversarial training recipe
- Our recipe generalizes to larger variants and different architectures
- One potential reason of why the recipe matters so much: it influences the inner part of adversarial training

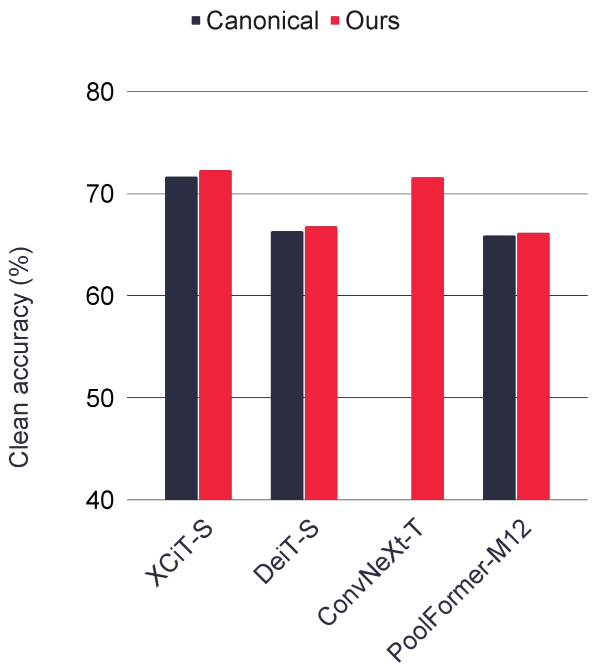
The optimal recipe scales up!

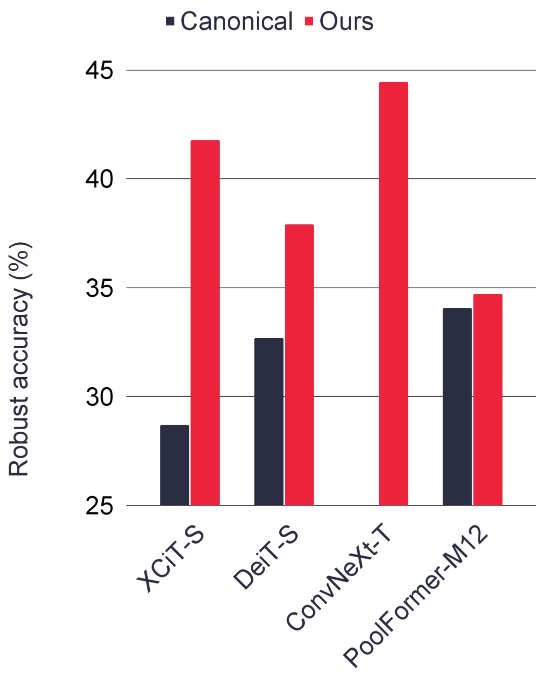


Comparison of our robust models to models from other works. The GELU ResNet-50 is from Bai et al. [2021] and the WRN-50-2 is from Salman et al. [2020].

[10] Bai, Yutong, et al. "Are Transformers more robust than CNNs?." Advances in Neural Information Processing Systems 34 (2021).
[11] Salman, Hadi, et al. "Do adversarially robust imagenet models transfer better?." Advances in Neural Information Processing Systems 33 (2020): 3533-3545.

And generalizes to other architectures!





Our recipe brings significant improvements for a range of architectures.

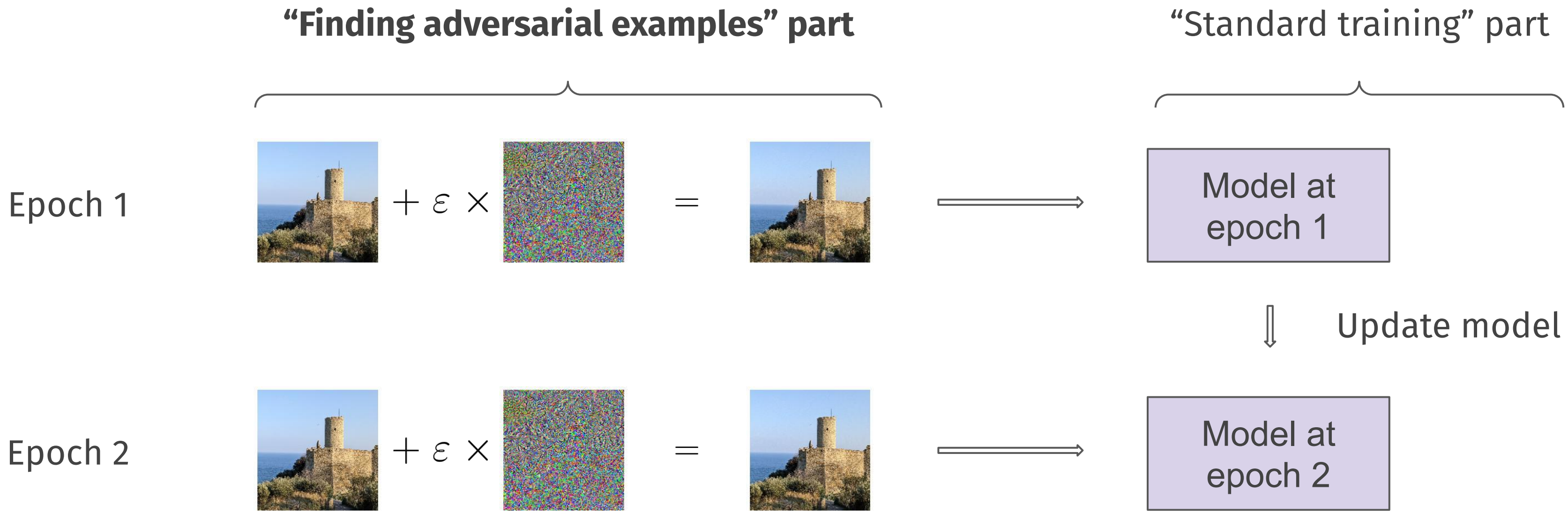
Contributions of this work

- Vision Transformers can be competitive at adversarial training, but need a custom adversarial training recipe
- Our recipe generalizes to larger variants and different architectures
- One potential reason of why the recipe matters so much: it influences the inner part of adversarial training

Contributions of this work

- Vision Transformers can be competitive at adversarial training, but need a custom adversarial training recipe
- Our recipe generalizes to larger variants and different architectures
- One potential reason of why the recipe matters so much: it influences the inner part of adversarial training

The recipe influences adversarial training's inner loop



When training, we want to generate strong adversarial examples with few PGD steps.

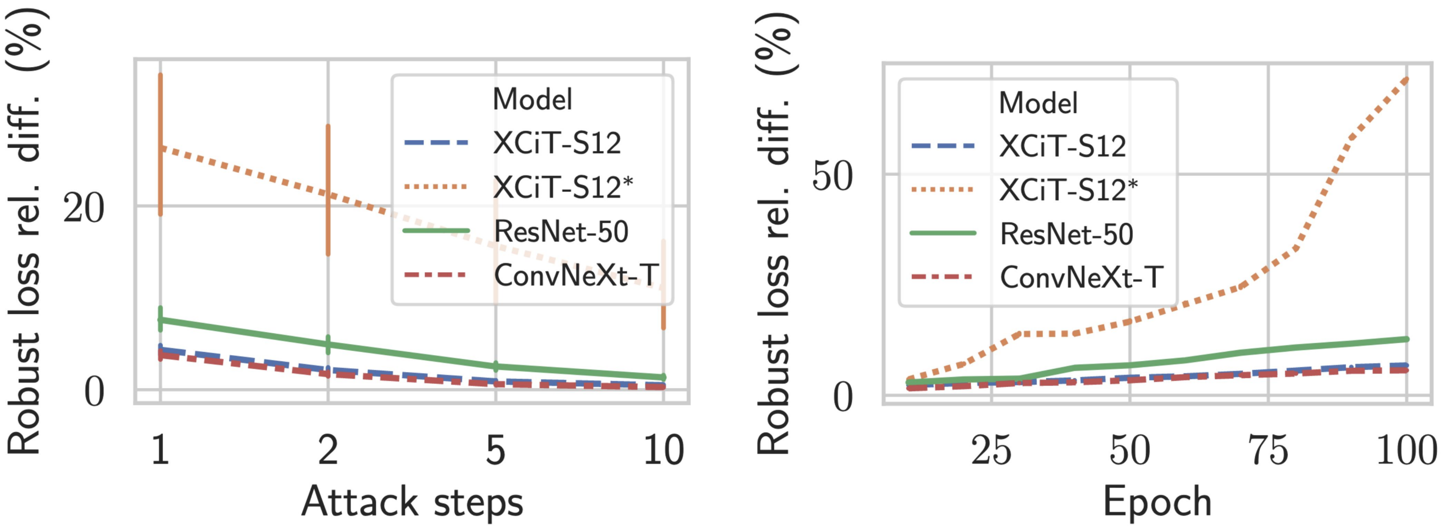
[16] Xie, Cihang, et al. "Smooth adversarial training." arXiv preprint arXiv:2006.14536 (2020).

The recipe influences adversarial training's inner loop

$$d_k = \frac{\mathcal{L}(\mathbf{x} + \boldsymbol{\delta}_k, \mathrm{y}; \boldsymbol{\theta}) - \mathcal{L}(\mathbf{x} + \boldsymbol{\delta}_O, \mathrm{y}; \boldsymbol{\theta})}{\mathcal{L}(\mathbf{x} + \boldsymbol{\delta}_O, \mathrm{y}; \boldsymbol{\theta})}$$

A small relative difference suggests that we need few PGD steps to get a strong enough adversarial example.

The recipe influences adversarial training's inner loop



* canonical recipe

- Models that end up being more robust show smaller relative differences throughout the training, at different relative steps.
- The relative differences for XCiT-S12 trained with the canonical recipe are significantly larger!

This work

- Vision Transformers can be competitive at adversarial training, but need a custom adversarial training recipe
- Our recipe generalizes to larger variants and different architectures
- One potential reason of why the recipe matters so much: it influences the inner part of adversarial training

Questions?

edebenedetti@inf.ethz.ch
@edoardo_debe

Paper Code

Backup slides

ViTs and variations — Vision Transformer

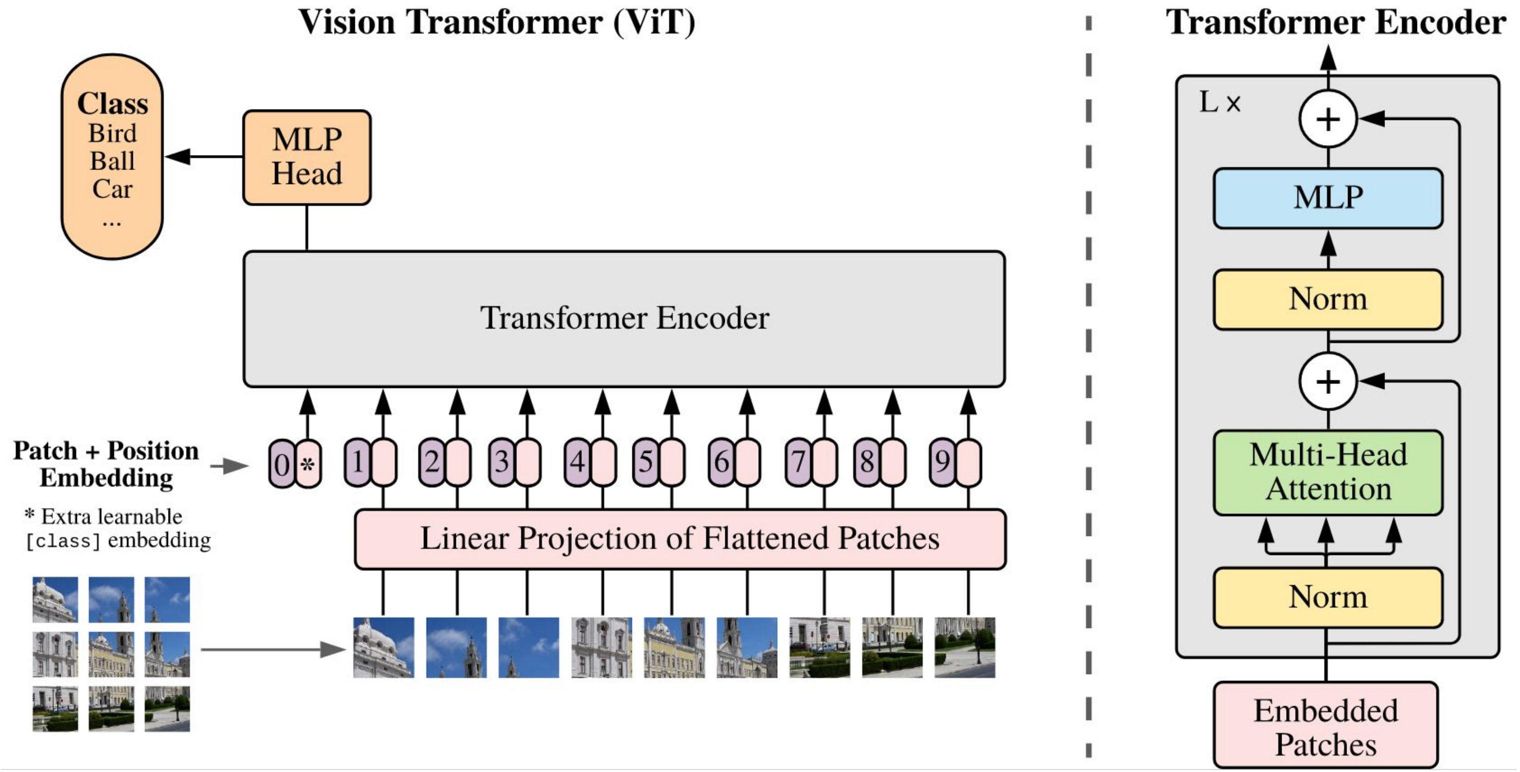


Image from Dosovitskiy et al. [2021]

ViTs and variations — Class Attention

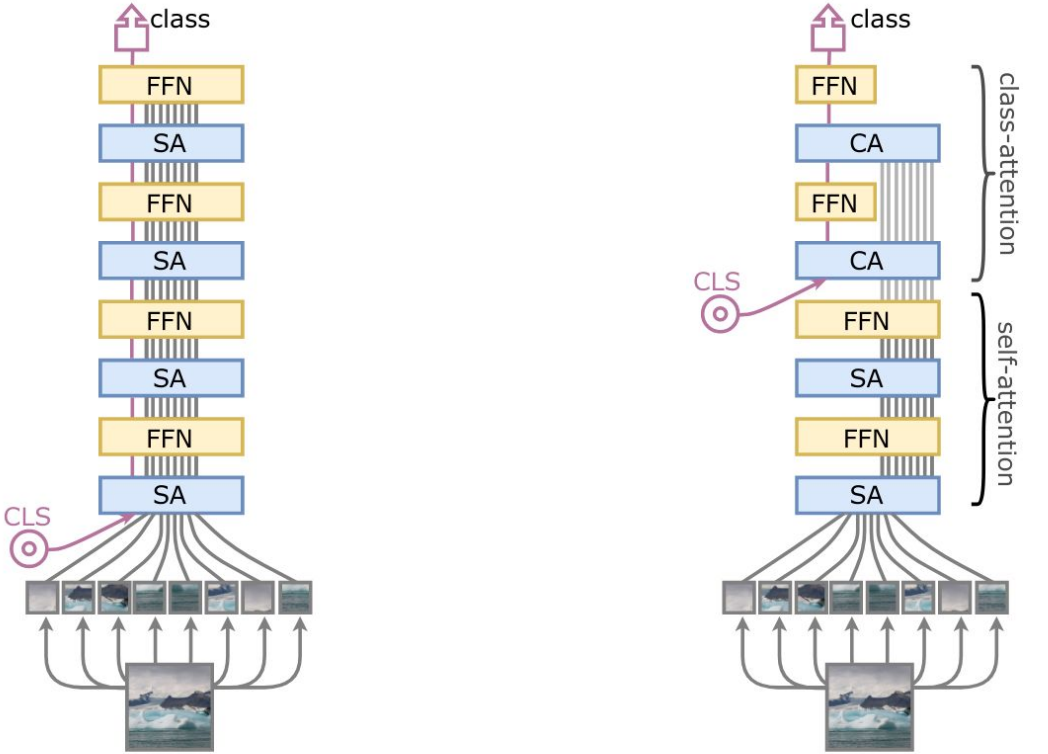


Image from Touvron et al. [2021]

ViTs and variations — Cross-covariance ViT (XCiT)

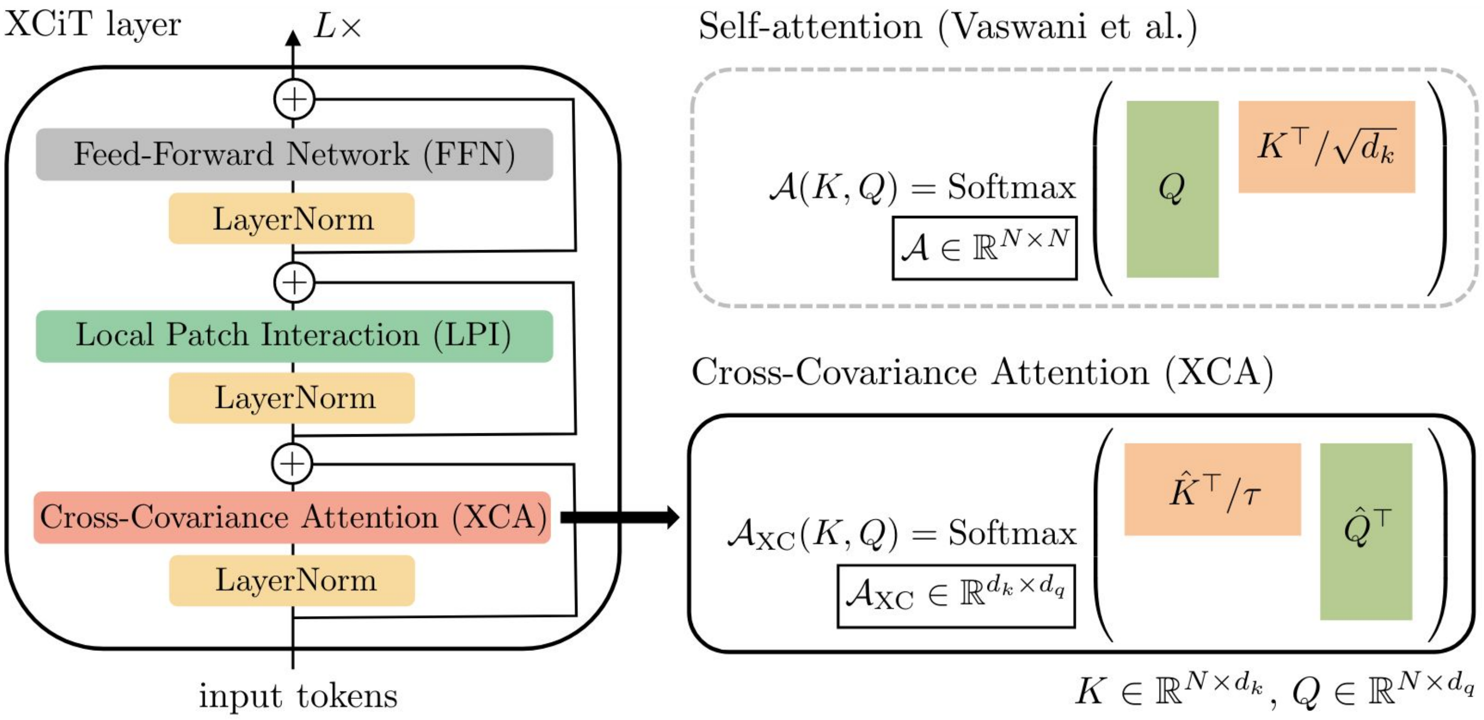
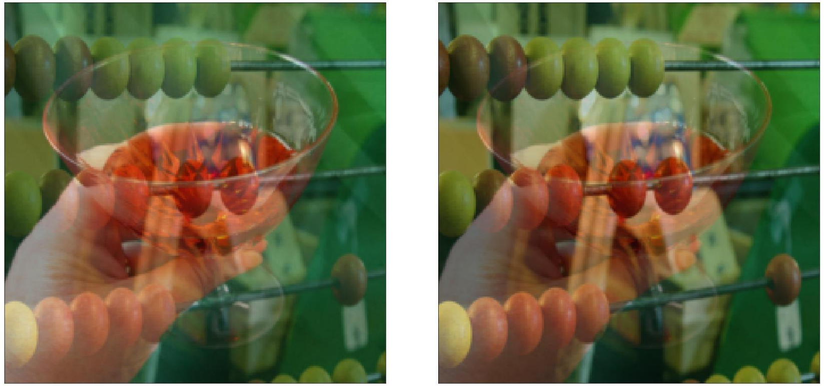


Image from El-Nouby et al. [2021]

Data augmentations



MixUp

CutMix

Zhang, Hongyi, et al. "mixup: Beyond empirical risk minimization." arXiv preprint arXiv:1710.09412 (2017).
Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. ICCV

Data augmentations

RandAugment

Random Erasing

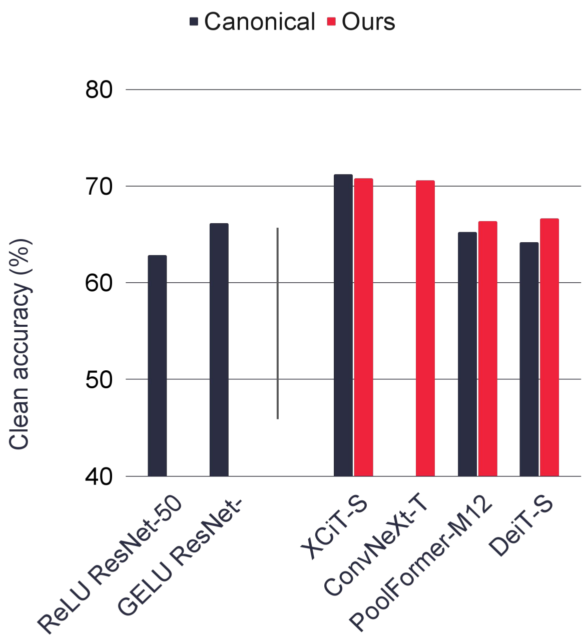
Zhong, Zhun, et al. "Random erasing data augmentation." Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 07. 2020.
Cubuk, Ekin D., et al. "Randaugment: Practical automated data augmentation with a reduced search space." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020.

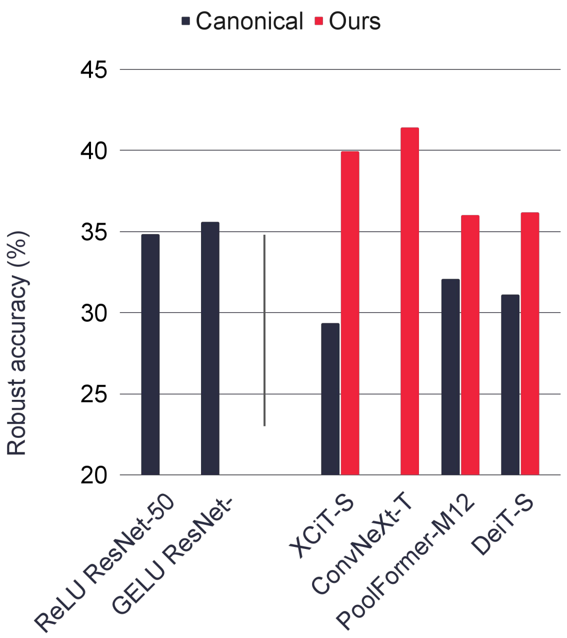
Set-up for standard training ≠ set-up of adversarial training

Data Augmentation Policy				Standard training	Adversarial training	
MixUp	CutMix	RandAugment	Random Erasing	*Clean*	*Clean*	*APGD-CE*
✗	✗	✗	✓	77.22	**67.28**	$\boxed{\textbf{39.22}}$
✗	✗	✗	✗	76.60	66.78	**39.22**
✓	✗	✗	✗	76.34	61.04	38.56
✓	✗	✗	✓	76.02	60.46	38.26
✓	✓	✗	✗	76.48	62.04	38.18
✗	✗	✓	✗	$\boxed{\textbf{78.62}}$	65.34	$\boxed{37.64}$
✗	✗	✓	✓	78.08	64.76	37.62
✓	✓	✓	✓	75.32	56.64	$\boxed{35.38}$

Top performing data augmentation set-ups for both standard and adversarial training. The tuned set-up improves the original one by 3.84%.

Our recipe generalizes to the L2 threat-model

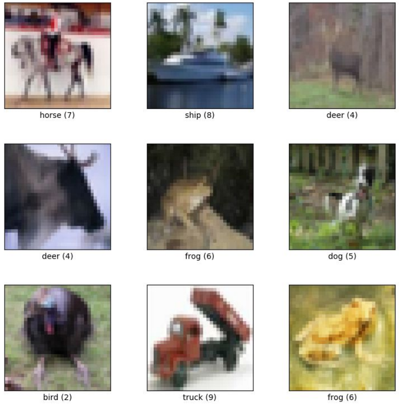


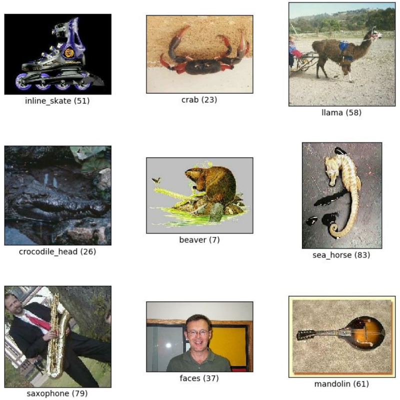


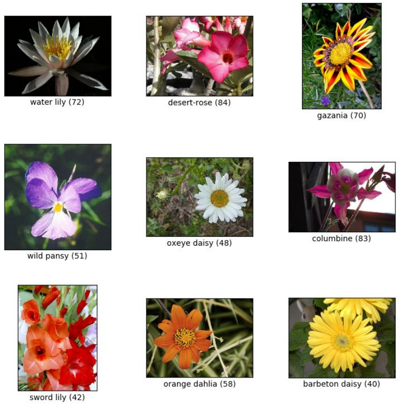
Comparison between the canonical recipe and our recipe on ImageNet for L2 perturbations with ε = 3.0. The ReLU ResNet-50 is from Salman et al. [2020].

Pre-training and model adaptation

- We pre-train XCiT-S on ImageNet for ε = 8/255
- We adapt the patch embedding layer to fine-tune on CIFAR-10 and CIFAR-100 which have 32x32 resolution (vs. ImageNet's 224x224)
- We fine-tune on CIFAR-10, CIFAR-100, Caltech-101, and Oxford Flowers



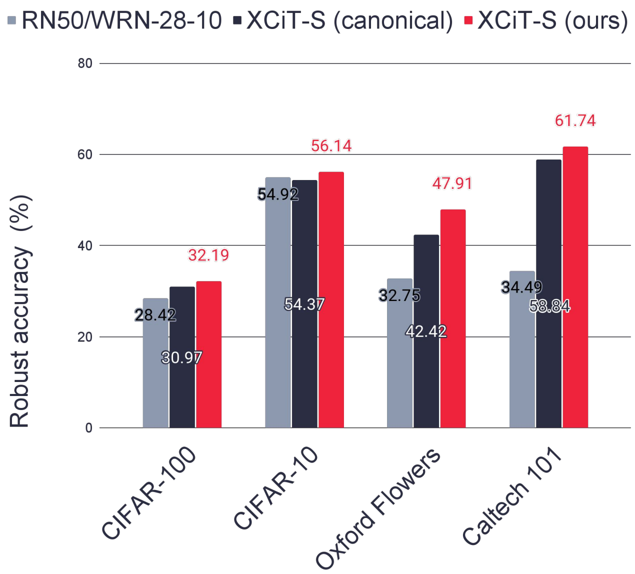




Dataset samples from the TensorFlow Datasets website

[12] Learning Multiple Layers of Features from Tiny Images, Alex Krizhevsky, 2009
[13] Fei-Fei, Li, Rob Fergus, and Pietro Perona. "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories." 2004 conference on computer vision and pattern recognition workshop. IEEE, 2004.
[14] Nilsback, Maria-Elena, and Andrew Zisserman. "Automated flower classification over a large number of classes." 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. IEEE, 2008.

Pre-training and model adaptation



Comparison between XCiT-S12 trained with our recipe vs. the canonical recipe and WideResNet-28-10 from Hendrycks et al. [2019] (for CIFAR) / ResNet-50 from Salman et al. [2020] (for Oxford Flowers and Caltech-101)

[15] Hendrycks, Dan, Kimin Lee, and Mantas Mazeika. "Using pre-training can improve model robustness and uncertainty." International Conference on Machine Learning. PMLR, 2019.

Model adaptation

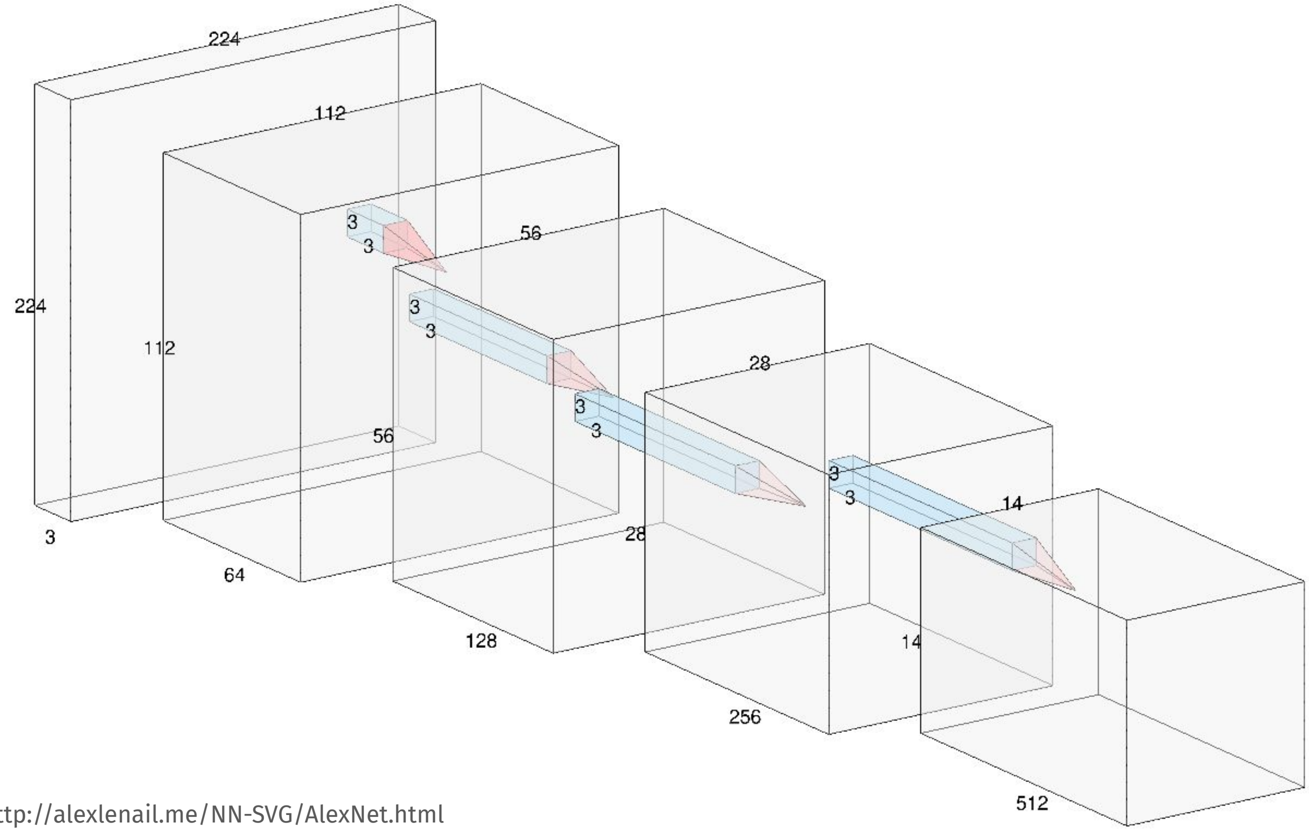


Diagram generated with http://alexlenail.me/NN-SVG/AlexNet.html

Model adaptation

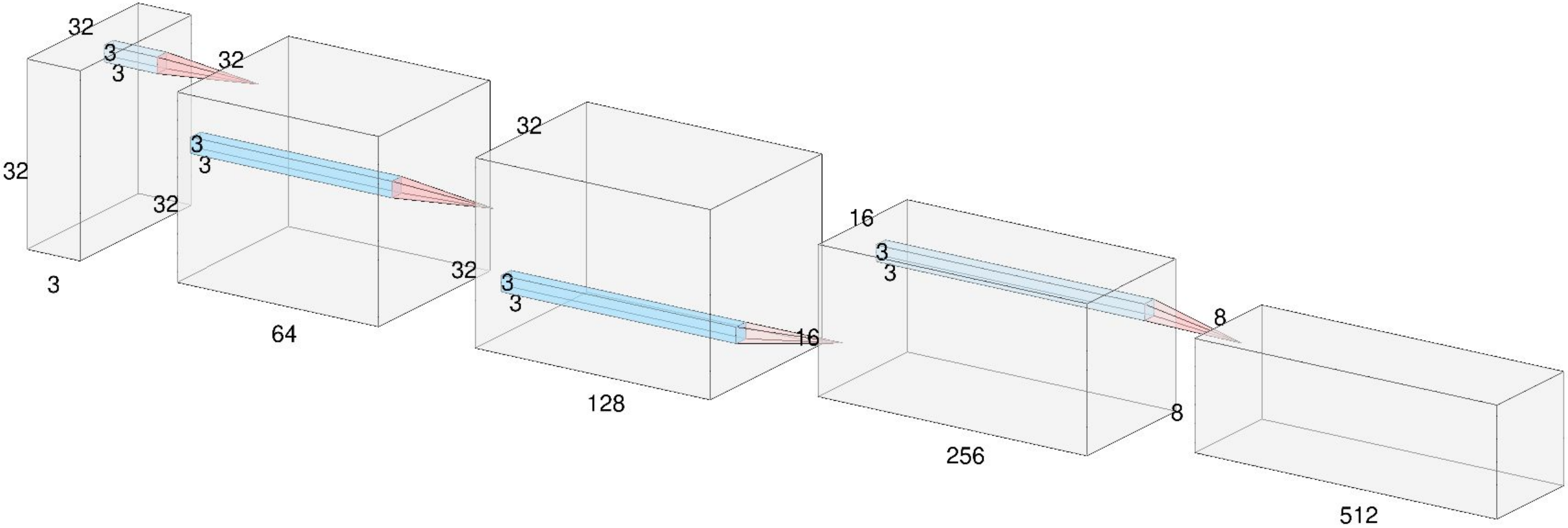


Diagram generated with http://alexlenail.me/NN-SVG/AlexNet.html

CIFAR-10

(b) **CIFAR-10 adversarial fine-tuning.**

Model	Clean Accuracy	AA Accuracy
WideResNet-28-10 [59]	87.11	54.92
ResNet-50	84.80	41.56
XCiT-S12 *(c)*	89.07	54.37
XCiT-S12 *(ours)*	90.06	56.14
XCiT-M12 *(ours)*	91.30	57.27
XCiT-L12 *(ours)*	**91.73**	**57.58**

CIFAR-100

(d) **CIFAR-100 adversarial fine-tuning.**

Model	Clean Accuracy	AA Accuracy
WideResNet-28-10 [59]	59.23	28.42
ResNet-50	61.28	22.01
XCiT-S12 *(c)*	65.44	30.97
XCiT-S12 *(ours)*	67.34	32.19
XCiT-M12 *(ours)*	69.21	34.21
XCiT-L12 *(ours)*	**70.76**	**35.08**

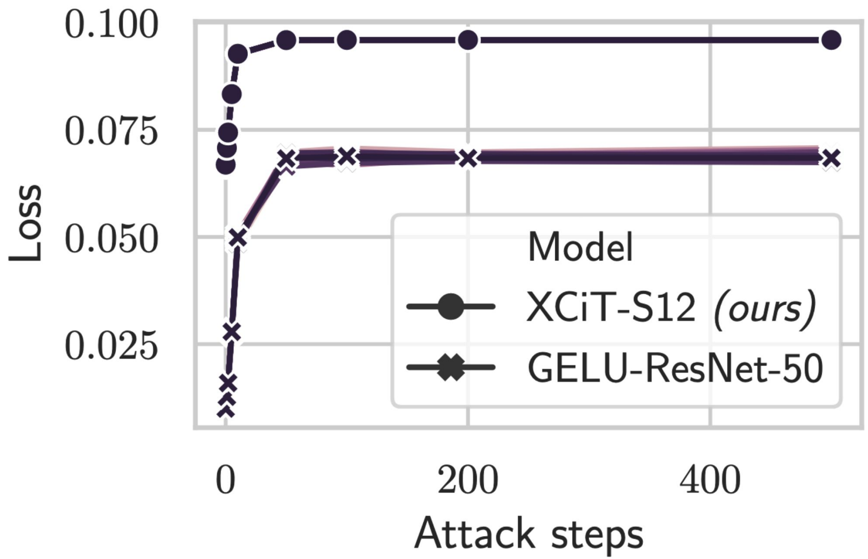
CIFAR-100

Leaderboard: CIFAR-100, $\ell_\infty = 8/255$, untargeted attack

Show 15 entries   Search: Papers, architectures, ve

Rank	Method	Standard accuracy	AutoAttack robust accuracy	Best known robust accuracy	AA eval. potentially unreliable	Extra data	Architecture	Venue
1	Uncovering the Limits of Adversarial Training against Norm-Bounded Adversarial Examples	69.15%	36.88%	36.88%	×	☑	WideResNet-70-16	arXiv, Oct 2020
2	A Light Recipe to Train Robust Vision Transformers	70.76%	35.08%	35.08%	×	☑	XCiT-L12	arXiv, Sep 2022
3	Fixing Data Augmentation to Improve Adversarial Robustness *It uses additional 1M synthetic images in training.*	63.56%	34.64%	34.64%	×	×	WideResNet-70-16	arXiv, Mar 2021
4	A Light Recipe to Train Robust Vision Transformers	69.21%	34.21%	34.21%	×	☑	XCiT-M12	arXiv, Sep 2022
5	Robustness and Accuracy Could Be Reconcilable by (Proper) Definition *It uses additional 1M synthetic images in training.*	65.56%	33.05%	33.05%	×	×	WideResNet-70-16	ICML 2022
6	A Light Recipe to Train Robust Vision Transformers	67.34%	32.19%	32.19%	×	☑	XCiT-S12	arXiv, Sep 2022

Is PGD-200 a good oracle?



Saturating of the cross-entropy loss in separate runs of PGD attacks with different numbers of steps, perturbing the same input.

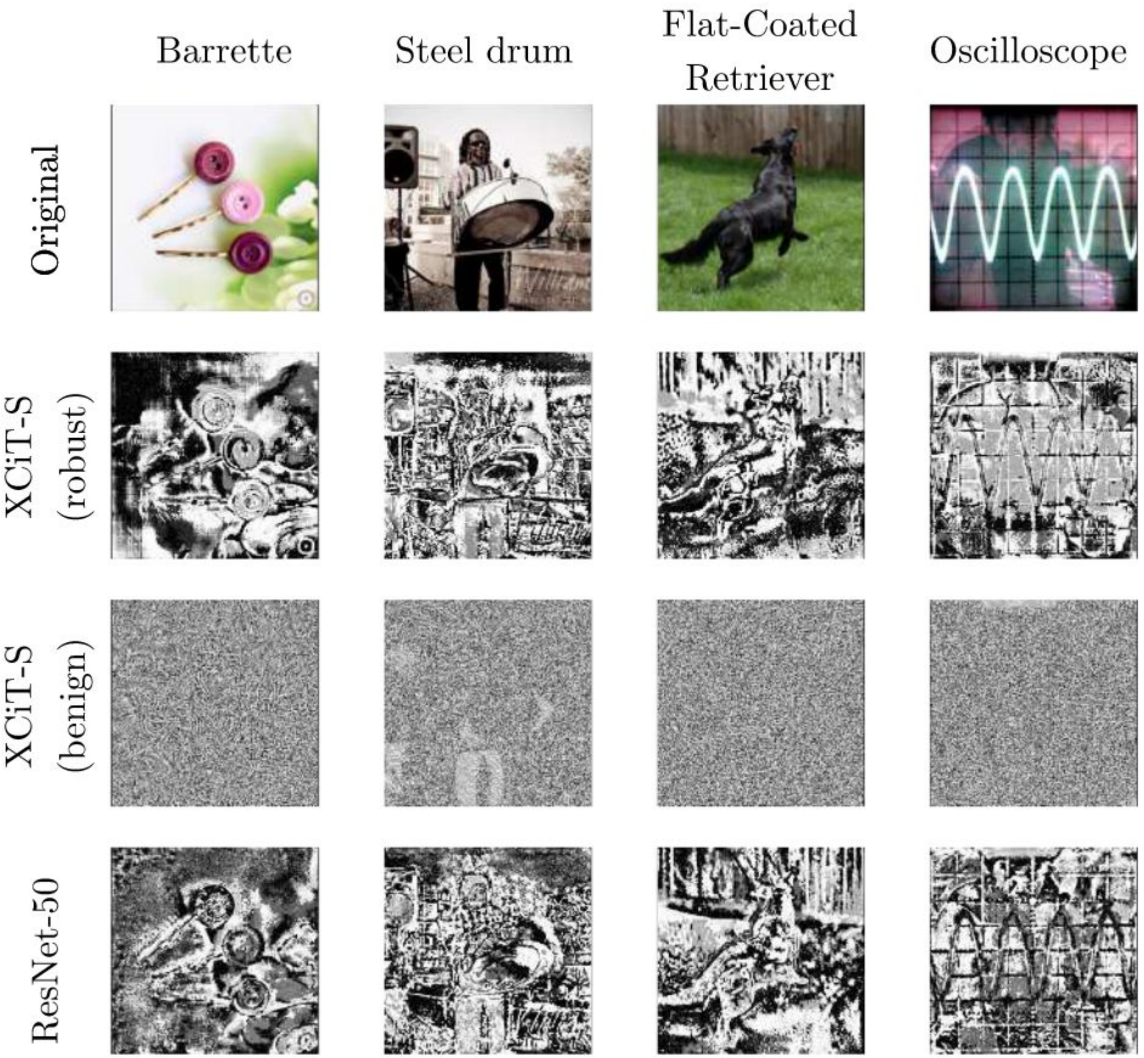
XCiT's attacks are more perceptual

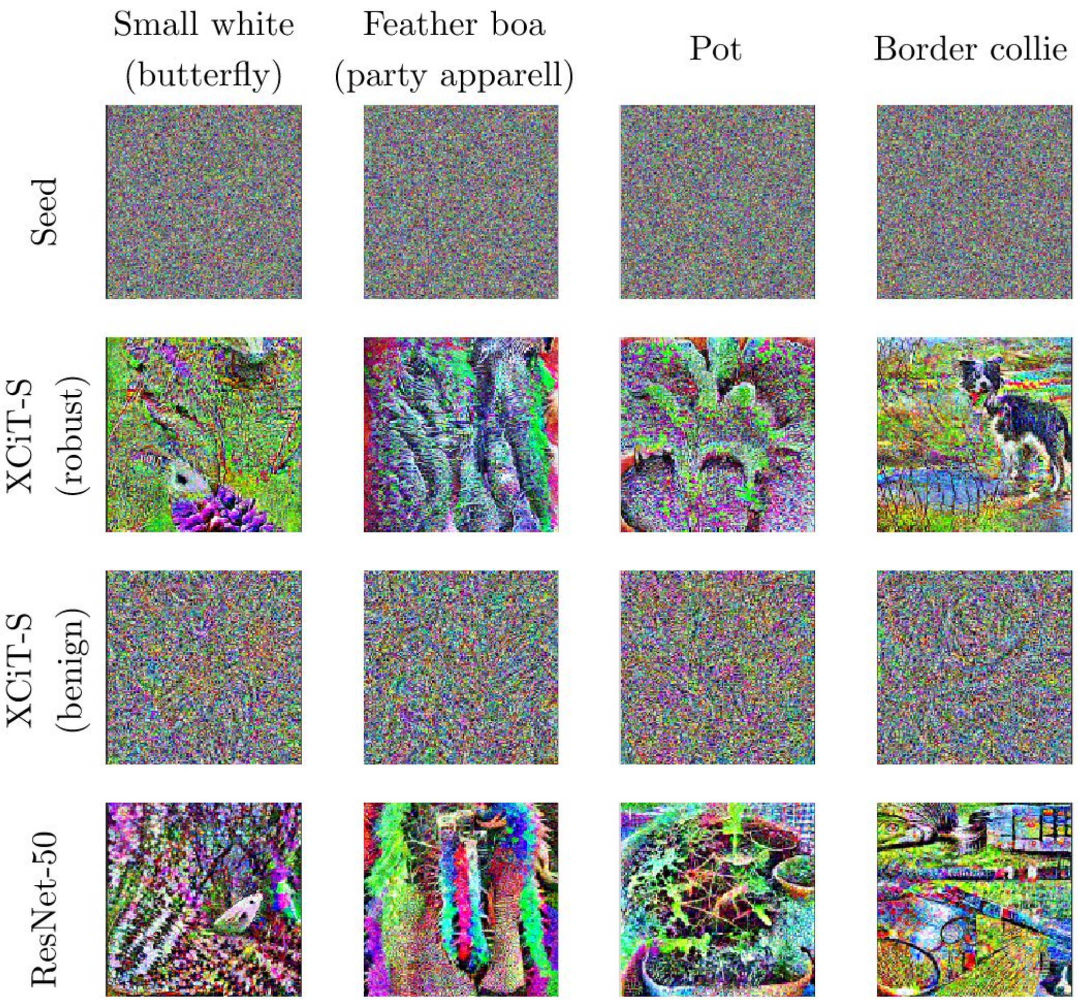
- We rescale the perturbations in *[0, 1]*
- We classify the perturbations using state of the art ImageNet models
- More perceptual perturbations should be classified more correctly

Perturbations generator		Classifier		
		ConvNeXt-XL	BeiT-L	Swin-L
Robust	XCiT-S12	43.86	49.52	40.24
	ResNet-50	38.40	45.02	36.70
Non-robust	XCiT-S12	0.84	0.78	0.84
	ResNet-50	0.82	0.74	0.80

Robust ResNet from Bai et al. [2021], non-robust ResNet from Wightman et al. [2021], non-robust XCiT from El-Nouby et al. [2021]

[22] Wightman, Ross, Hugo Touvron, and Hervé Jégou. "Resnet strikes back: An improved training procedure in timm." arXiv preprint arXiv:2110.00476 (2021).
[23] Ali, Alaaeldin, et al. "Xcit: Cross-covariance image transformers." Advances in neural information processing systems 34 (2021).


Barrette
Steel drum
Flat-Coated Retriever
Oscilloscope
Original
XCiT-S (robust)
XCiT-S (benign)
ResNet-50


Small white (butterfly)
Feather boa (party apparell)
Pot
Border collie
Seed
XCiT-S (robust)
XCiT-S (benign)
ResNet-50