Boosting Model Robustness by Leveraging Data Augmentations, Stability Training, and Noise Injections

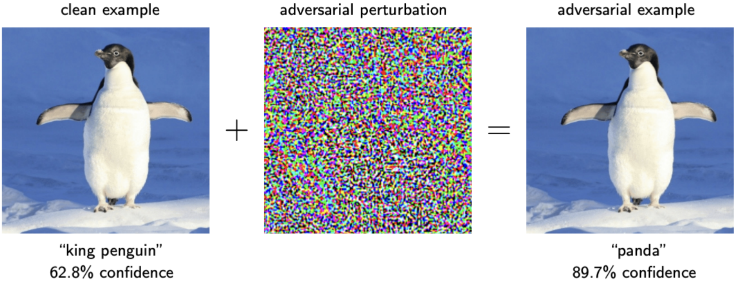
Soon Hoe Lim (Nordita, KTH & Stockholm U)

Joint work with N. Benjamin Erichson (U of Pittsburgh), Francisco Utrera (ICSI & U of Pittsburgh), Winnie Xu (U of Toronto), Ziang Cao (U of Pittsburgh), and Michael Mahoney (ICSI & UC Berkeley)

Applied Machine Learning Days 2022

Model Robustness Matters!

- Deep learning models are typically brittle and sensitive to noisy and adversarial environments
- For many real-world applications, obtaining stable and robust statistical performance is more important than simply achieving SOTA predictive performance
- Here we focus on input stability (robustness) with respect to common data corruptions and domain shifts that naturally occur in many real-world applications



- Szegedy et al. "Intriguing properties of neural networks." ICLR (2014).
- Goodfellow et al. "Explaining and harnessing adversarial examples." ICLR (2015).

NoisyMix

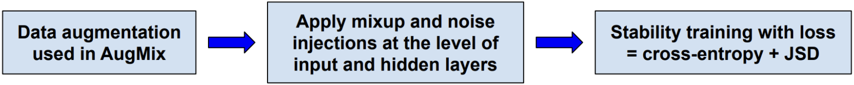
Four common methods to improve model robustness to input perturbations are:

- Data augmentations
- Stability training
- Mixup
- Noise injections

How can we leverage the strength of these methods to further improve both model robustness and test accuracy?

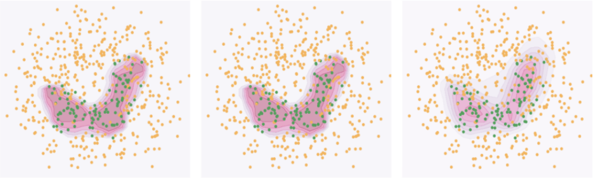
We introduce NoisyMix, a training scheme that judiciously combines all of the above components in a single setup to boost both robustness and accuracy in classification tasks.

Sounds simple, but the devil is in the details to get the method right:

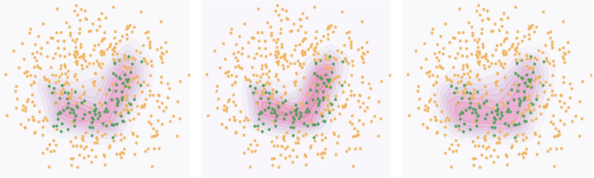


Diving Deeper into NoisyMix

The advantage of NoisyMix compared to other schemes is illustrated on a binary classification task on a noisy toy dataset (without augmentation), where it can be seen that NoisyMix is most effective at smoothing the decision boundary and yields the best test accuracy:



Baseline (86.8%). Noise injection (86.8%). Manifold Mixup (87.4%).

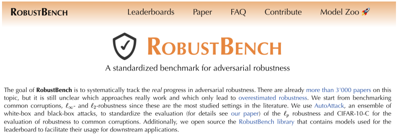


NFM (87.6%). Manifold Mixup + JSD (88.0 %). *NoisyMix* (88.8%).

Moreover, we provide theory to understand the effects of NoisyMix through the lens of implicit regularization and show that minimizing the NoisyMix loss can lead to a small regularized adversarial loss and a stable model (see paper).

Empirical Results

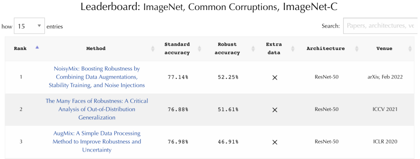
We benchmark common corruptions with **RobustBench**, and find that NoisyMix (currently) tops the leaderboards (CIFAR-10-C, CIFAR-100-C, and ImageNet-C) there.


ROBUSTBENCH Leaderboards Paper FAQ Contribute Model Zoo

ROBUSTBENCH

A standardized benchmark for adversarial robustness

The goal of **RobustBench** is to systematically track the *real* progress in adversarial robustness. There are already more than 3'000 papers on this topic, but it is still unclear which approaches really work and which only lead to overestimated robustness. We start from benchmarking common corruptions, ℓ_∞- and ℓ_2-robustness since these are the most studied settings in the literature. We use AutoAttack, an ensemble of white-box and black-box attacks, to standardize the evaluation (for details see our paper) of the ℓ_p robustness and CIFAR-10-C for the evaluation of robustness to common corruptions. Additionally, we open source the RobustBench library that contains models used for the leaderboard to facilitate their usage for downstream applications.


Leaderboard: ImageNet, Common Corruptions, ImageNet-C

how 15 entries Search: Papers, architectures, ve

Rank	Method	Standard accuracy	Robust accuracy	Extra data	Architecture	Venue
1	NoisyMix: Boosting Robustness by Combining Data Augmentations, Stability Training, and Noise Injections	77.14%	52.25%	×	ResNet-50	arXiv, Feb 2022
2	The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization	76.88%	51.61%	×	ResNet-50	ICCV 2021
3	AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty	76.98%	46.91%	×	ResNet-50	ICLR 2020

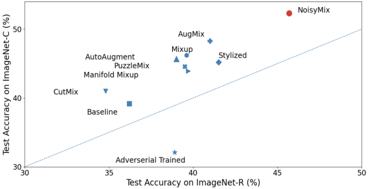


Table 1: Clean test accuracy of ResNet-50 models on ImageNet, and average robust accuracy on ImageNet-C and ImageNet-R (higher values are better). In addition, we show the mean flip rate for ImageNet-P (lower values are better) and the robustness to adversarial examples constructed with FGSM (higher values are better). The values in parenthesis are the average robust accuracy for ImageNet-C and mean flip probability for ImageNet-P excluding noise perturbations.

	ImageNet (↑%)	ImageNet-C (↑%)	ImageNet-R (↑%)	ImageNet-P (↓%)	FGSM (↑%)
Baseline (He et al., 2016a)	76.1	39.2 (42.3)	36.2	58.0 (57.8)	6.6
Adversarial Trained (Salman et al., 2020)	63.9	32.1 (35.4)	38.9	33.3 (33.4)	**43.1**
Stylized ImageNet (Geirhos et al., 2018)	74.9	45.2 (46.6)	41.5	54.4 (55.2)	7.8
AutoAugment (Cubuk et al., 2019)	77.6	45.7 (47.3)	39.0	56.5 (57.7)	9.9
Mixup (Zhang et al., 2018)	77.5	46.2 (48.4)	39.6	56.4 (58.7)	23.5
Manifold Mixup (Verma et al., 2019)	76.7	43.9 (46.5)	39.7	56.0 (58.2)	27.7
CutMix (Yun et al., 2019)	78.6	41.0 (43.1)	34.8	58.6 (59.9)	33.7
Puzzle Mix (Kim et al., 2020)	**78.7**	44.6 (46.4)	39.5	55.5 (57.0)	28.4
AugMix (Hendrycks et al., 2020)	77.5	48.3 (50.5)	41.0	37.6 (37.2)	9.9
NoisyMix (ours)	77.6	**52.3 (52.4)**	**45.7**	**28.5 (29.7)**	29.6

More results are available at:

- Paper: `https://arxiv.org/abs/2202.01263`
- RobustBench: `https://robustbench.github.io/`