Revisiting Adversarial Training under Long-Tailed Distributions

Xinli Yue Ningping Mou Qian Wang Lingchen Zhao

School of Cyber Science and Engineering, Wuhan University

Overview

- We discover that BSL is the most critical component of RoBal, and the streamlined method AT-BSL can improve the efficiency of RoBal.
- We observe that data augmentation substantially mitigates robust overfitting and improves robustness under long-tailed distributions.
- We propose a hypothesis about how data augmentation improves robustness and validate this hypothesis through experiments.
- Comprehensive empirical evidence demonstrates that our discoveries generalize across multiple common scenarios.

Adversarial Training

The insight of adversarial training is integrating adversarial examples into the training set, thereby improving the generalizability of the model to such examples.

$$\underset{\theta_m}{\operatorname{argmin}} \mathcal{L}_{\min}(\theta_m; x', y), \text{where } x' = \underset{\|x'-x\|_p \le \epsilon}{\operatorname{argmax}} \mathcal{L}_{\max}(\theta_m; x', y).$$

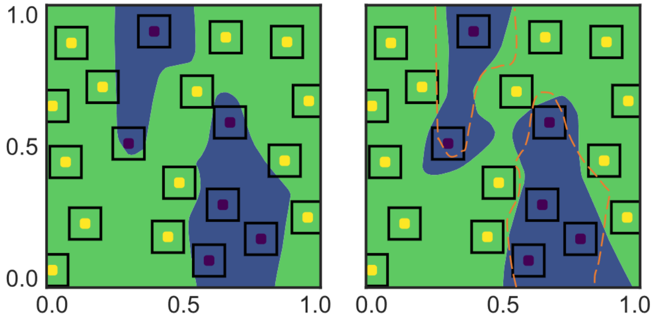


Figure from (Zhang et al., 2019)

- Zhang et al., Theoretically Principled Trade-off between Robustness and Accuracy, ICML 2019.

Long-Tailed Distributions

Long-tailed distributions refer to a common imbalance in the training set. Models trained under such distribution tend to exhibit a bias towards the head classes, resulting in poor performance for the tail classes.

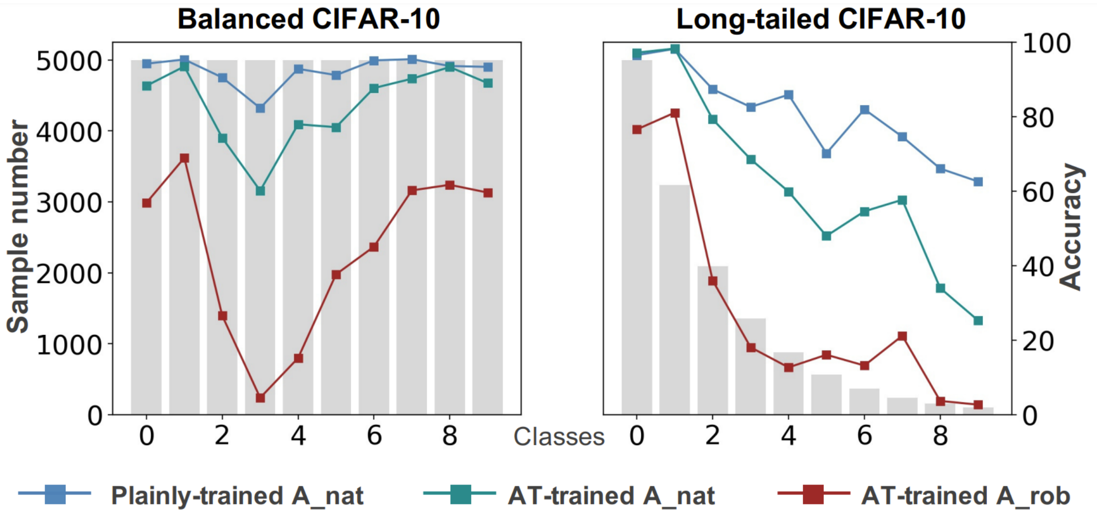


Figure from (Wu et al., 2021)

- Wu et al., Adversarial Robustness under Long-Tailed Distribution, CVPR 2021.

RoBal

Cosine Classifier

- RoBal employs a cosine classifier to minimize the scale effects of features and weights.

$$h(f(x))_i = s \cdot \left(\frac{W_i^T f(x)}{\|W_i\| \parallel f(x) \parallel} \right) + b_i = s \cdot \cos \theta_i + b_i.$$

Balanced Softmax Loss (BSL)

- An intuitive and widely adopted approach to address class imbalance is assigning class-specific biases during training for cross-entropy loss.

$$\mathcal{L}_0\big(h(f(x)), y\big) = -\log \left(\frac{e^{s \cdot \cos \theta_y + b_y}}{\sum_i \ e^{s \cdot \cos \theta_i + b_i}} \right).$$

- Wu et al., Adversarial Robustness under Long-Tailed Distribution, CVPR 2021.

RoBal

Class-Aware Margin

- To address that BSL may degrade the quality of discriminative representations, RoBal designs a class-aware margin term, which assigns a larger margin value to head classes as a form of compensation.

$$m_i = \frac{\tau_m}{s} \log \frac{n_i}{n_{\min}} + m_0.$$

TRADES Regularization

- RoBal incorporates a Kullback-Leibler (KL) regularization term following TRADES.

$$\mathcal{L}_{\min} = \mathcal{L}_1\big(h\big(f(x')\big), \mathrm{y}\big) + \beta \cdot \mathrm{KL}\big(h\big(f(x')\big), h(f(x))\big).$$

- Zhang et al., Theoretically principled trade-off between robustness and accuracy, ICML 2019.

Ablation Studies of RoBal

The contribution of each component

- AT augmented with BSL (AT-BSL) outperforms the vanilla AT in both clean accuracy and adversarial robustness.
- Subsequent components do not yield significant improvements in robustness against AA, yet substantially increase both training time and memory usage.

Method	Components				Accuracy						Efficiency	
	Cos	BSL	CM	TRADES	Clean	FGSM	PGD	CW	LSA	AA	Time (s)	Memory (MiB)
AT [30]					54.91	32.21	28.05	28.28	28.73	26.75	21.36	**946**
AT-BSL		✓			70.21	37.44	31.91	**31.45**	**32.25**	29.48	**21.00**	**946**
AT-BSL-Cos	✓	✓			**71.99**	39.41	34.73	30.27	29.94	28.43	22.39	**946**
AT-BSL-Cos-TRADES	✓	✓		✓	69.31	39.62	34.87	30.19	30.15	28.64	38.91	1722
RoBal [45]	✓	✓	✓	✓	70.34	**40.50**	**35.93**	31.05	31.10	**29.54**	39.03	1722

AT-BSL

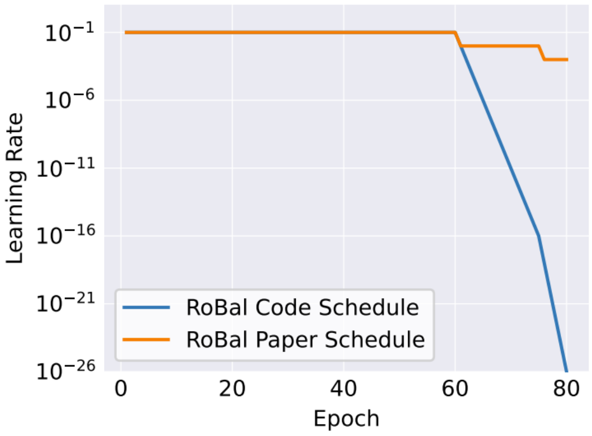
The advantages of AT-BSL

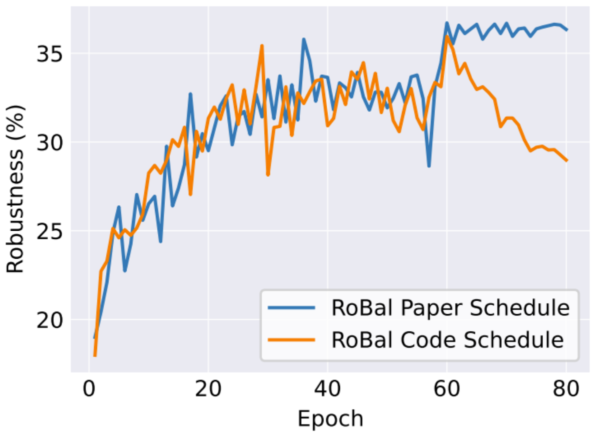
- AT-BSL in solation competes with the complete RoBal scheme in terms of clean accuracy and robustness against AA.
- AT-BSL significantly reduces the training time and GPU memory usage compared to RoBal.

$$\mathcal{L}_{\min} = \mathcal{L}_0\left(g\left(f\left(x'\right)\right), \mathrm{y}\right) = -\log\left(\frac{e^{z_y+b_y}}{\sum_i e^{z_i+b_i}}\right) = -\log\left(\frac{n_y^{\tau_b} \cdot e^{z_y}}{\sum_i n_i^{\tau_b} \cdot e^{z_i}}\right)$$

Robust Overfitting

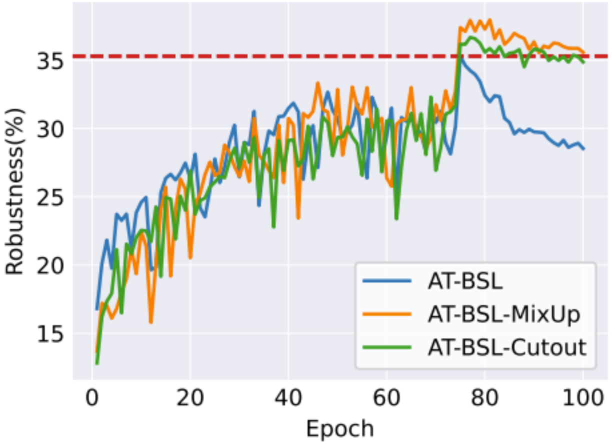
Adversarial training under long-tailed distributions also exhibits robust overfitting, similar to that under balanced distributions.

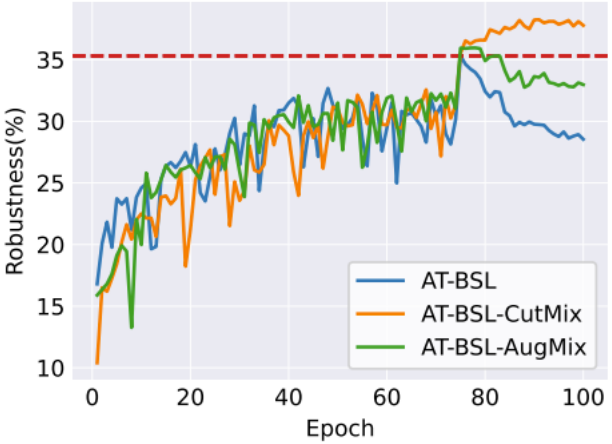




Data Augmentation

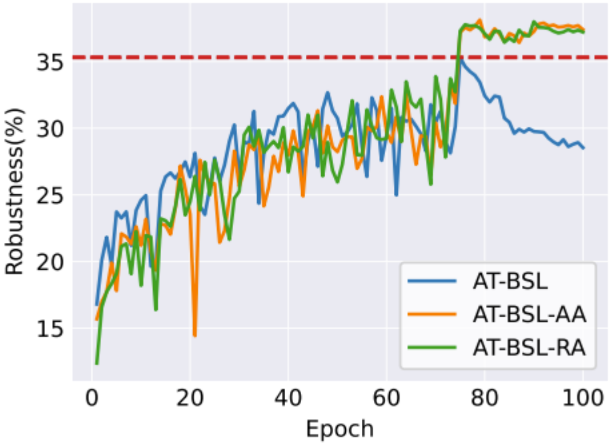
Data augmentation techniques like Mixup, Cutout, CutMix, AugMix can significantly alleviate robust overfitting.

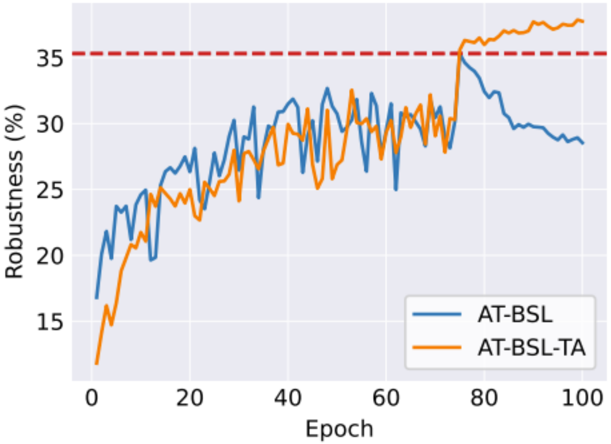




Unexpected Discoveries

The robustness achieved with each augmentation surpasses that of the vanilla AT-BSL, indicating that <u>data augmentation alone can indeed improve robustness</u>, which is inconsistent with conclusions drawn from balanced datasets.





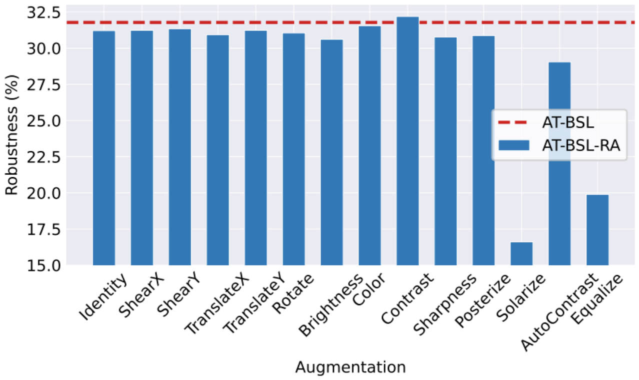
Hypothesis

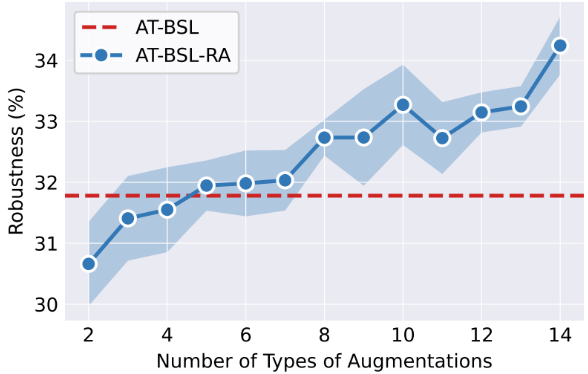
Formulating Hypothesis

- Data augmentation improves robustness by increasing the diversity of the training data, thus enabling models to learn richer representations.

Validating Hypothesis

- Single data augmentation cannot significantly improve robustness.
- Robustness consistently improves when more augmentations are added.





Evaluations

Robustness

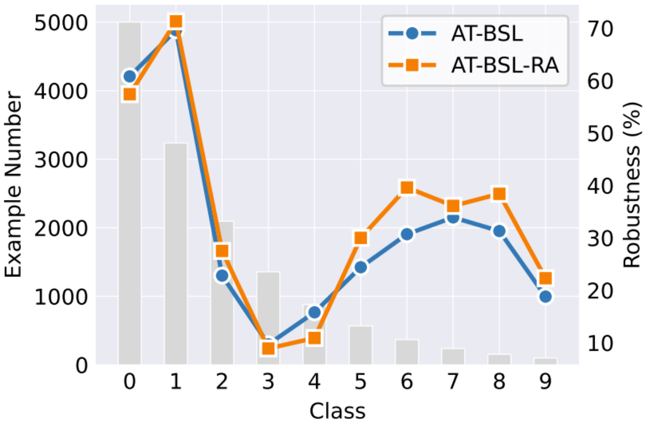
- AT-BSL with data augmentation obtains the highest clean accuracy and adversarial robustness.

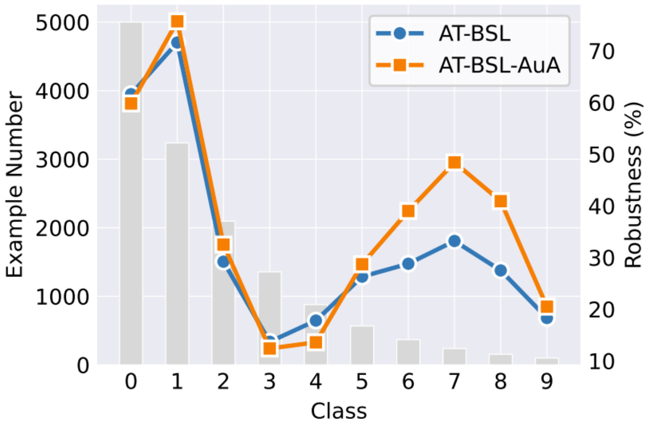
Method	Best Checkpoint						Last Checkpoint					
	Clean	FGSM	PGD	CW	LSA	AA	Clean	FGSM	PGD	CW	LSA	AA
AT [30]	59.21	31.88	27.88	28.19	29.81	27.07	58.25	29.77	25.29	25.71	29.83	24.94
TRADES [52]	51.28	31.58	28.70	28.45	28.36	27.72	53.85	30.44	26.23	26.57	26.77	25.59
MART [41]	49.13	34.33	32.32	30.73	30.13	29.60	52.48	33.95	31.09	29.64	29.43	28.67
AWP [44]	50.91	34.28	31.85	31.23	31.01	30.06	48.65	33.21	31.07	30.33	30.14	29.40
GAIRAT [53]	59.89	33.47	30.40	26.69	26.71	25.38	56.37	29.41	27.25	23.94	23.95	23.15
LAS-AT [19]	57.52	33.66	29.86	29.60	29.44	28.84	58.19	32.98	28.89	28.75	28.58	27.90
RoBal [45]	72.82	41.34	36.42	32.48	31.95	30.49	70.85	35.95	27.74	27.59	26.76	25.71
REAT [25]	73.16	41.32	35.94	35.28	35.67	33.20	67.76	34.51	27.75	28.17	31.82	26.66
AT-BSL	73.19	41.84	35.60	34.86	35.99	32.80	65.95	33.29	27.23	27.87	31.00	26.45
AT-BSL-AuA	**75.17**	**46.18**	**40.84**	**38.82**	**39.23**	**37.15**	**77.27**	**44.73**	**38.06**	**37.14**	**39.05**	**35.11**

Evaluations

Class-wise Robustness

- Apart from a few exceptions, data augmentation improves robustness across nearly all classes, particularly in the tail classes (classes 5 to 9).

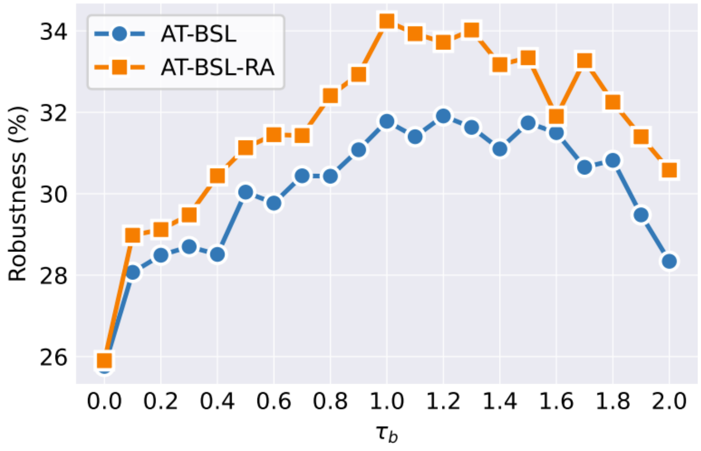


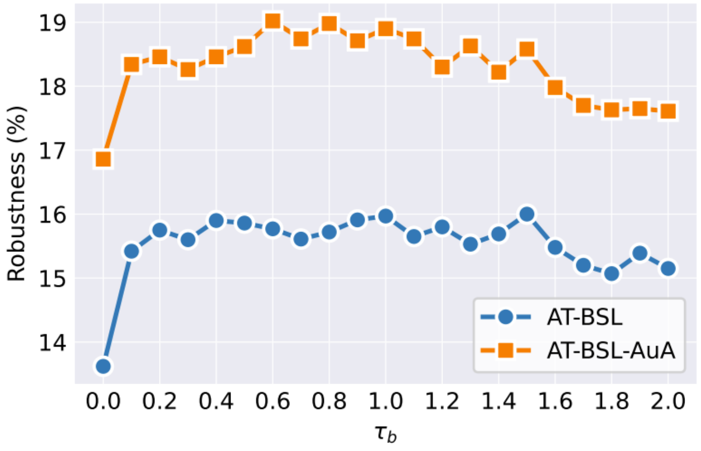


Futher Analysis

Effect of the Hyperparameter τ_b

- Including data augmentation strategies across all τ_b values consistently results in a significant robustness improvement compared to the vanilla AT-BSL.





Conclusion

- We first investigate the design of RoBal and identify Balanced Softmax Loss as the critical component.
- We discover that data augmentation not only mitigates robust overfitting but also improves robustness, and we validate the hypothesis we formulated.
- We conduct extensive experiments with various data augmentation strategies, model architectures, and datasets, affirming the generalizability of our findings.

Thanks!

lczhaocs@whu.edu.cn

https://github.com/NISPLab/AT-BSL