Rethinking Invariance Regularization in Adversarial Training to Improve Robustness-Accuracy Trade-off

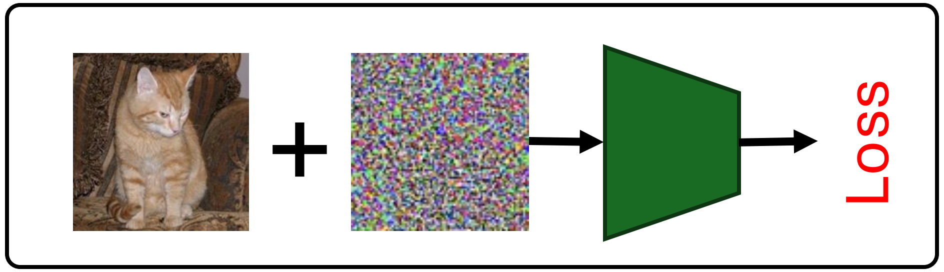
Futa Waseda [1],
Ching-Chun Chang[2],
Isao Echizen[1 2]

[1] The University of Tokyo, Japan
[2] National Institute of Information, Japan
Correspondence to: Futa Waseda <futa-waseda@g.ecc.u-tokyo.ac.jp>

Robustness accuracy trade-off

Adversarial training (AT) methods suffer from a trade-off between
- Clean accuracy (acc. on the clean samples)
- Robust accuracy (acc. on the adversarial examples)

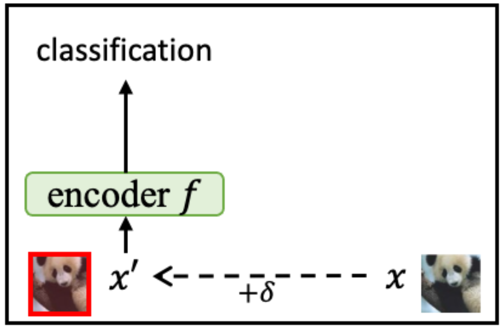
Adversarial Training (AT)



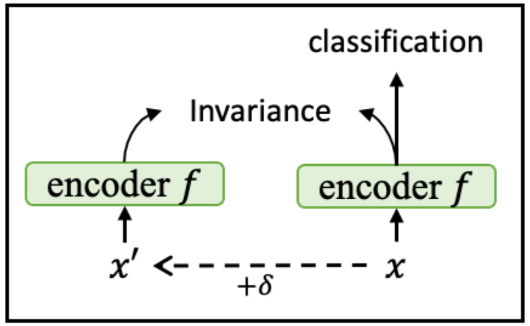
(CIFAR10,ResNet18)	Clean accuracy	Robust accuracy
Std. training	94.17	0.00
AT	82.45 ↓	50.04 ↑

This trade-off is a huge obstacle for real-world implementation of AT methods

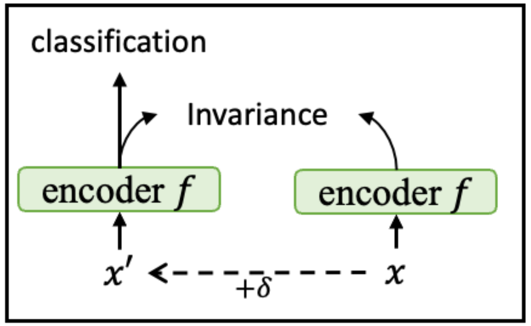
Invariance regularization-based Adversarial Training (AT) to mitigate robustness accuracy trade-off



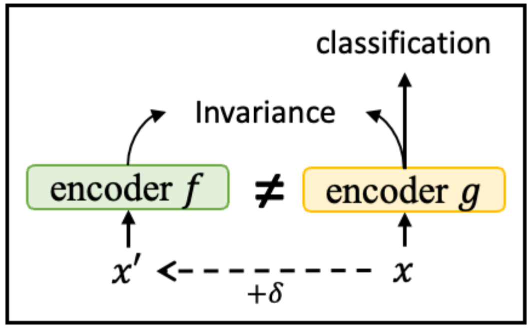
(a) AT [Madry+ 2018]



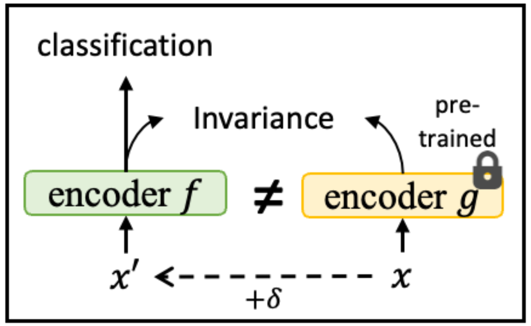
(b) TRADES [Zhang+ 2019]



(c) MART [Wang+ 2020]



(d) LBGAT [Cui+ 2021]

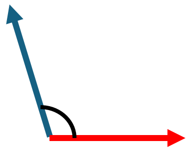


(e) ARREST [Suzuki+ 2023]

We closely analyze the challenges of using invariance regularization in adversarial training

Proposed Method: ARAT
(1) “Stop-gradient” to address gradient conflict

Issue 1 **Gradient Conflict**



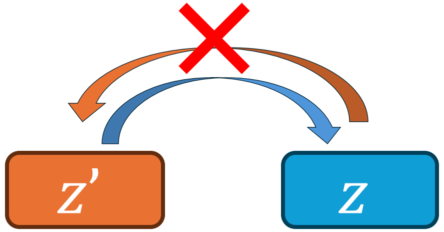
Naïve invariance regularization:

$$\mathcal{L}_{V0} = \underbrace{\alpha \cdot \mathcal{L}(f_\theta(x'), y) + \beta \cdot \mathcal{L}(f_\theta(x), y)}_{\text{Classification}} + \underbrace{\gamma \cdot Dist(z, z')}_{\text{Invariance}}$$

We observed that the gradients of the **Classification** and Invariance losses **conflicts**, leading to suboptimal convergence.

Solution 1 **Asymmetric invariance loss with “Stop-gradient” operation.**

$$Dist(z', z) = \left(\boxed{Dist(z', \mathrm{sg}(z))} + \cancel{\boxed{Dist(\mathrm{sg}(z'), z)}}\right)/2$$



The source of conflict! This corrupts representations.
We remove this term by applying a stop-gradient (sg) operation.

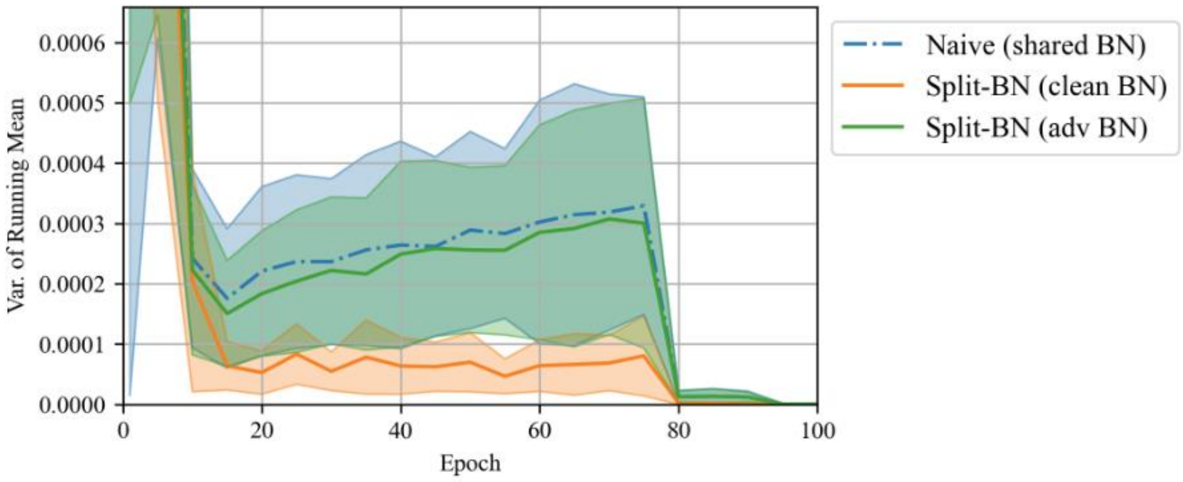
Proposed Method: ARAT
(2) Split-BN to address mixture distribution problem

Issue 2 **Mixture distribution problem**

The model struggles to handle both adversarial and clean inputs, due to **large distribution gap.**

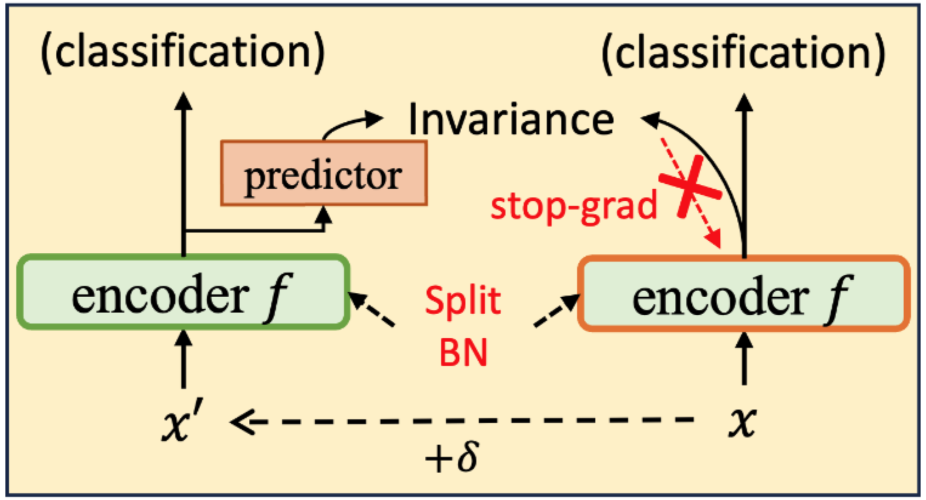
Solution 2 **Use separate batch normalizations (Split-BN) for adv. and clean inputs.**

With Split-BN, the model can properly handle both adversarial and clean inputs by normalizing them with different batch statistics, **improving training stability.**



ARAT effectively addresses the issues improving robustness-accuracy trade-off

- ARAT resolves
  - (1) **Gradient conflict** with **Stop-grad**
  - (2) **Mixture distribution problem** with **Split-BN**



ARAT (ours)

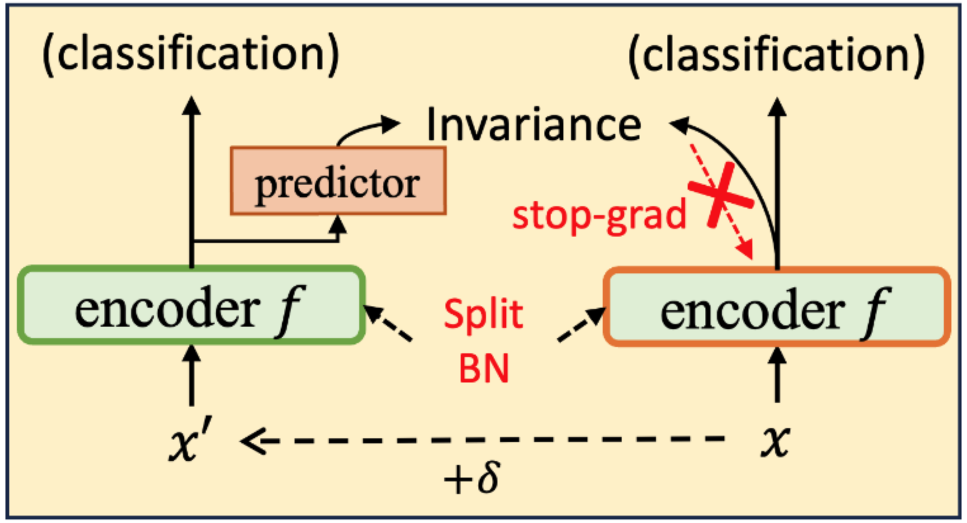
Robust acc.

		Stop-grad	Split-BN	Clean	AA	Grad-sim.
CIFAR10	ResNet-18			82.93	46.50	0.06
		✓		82.47	45.21	0.68
			✓	84.35	47.96	0.14
		✓	✓	**85.51**	**49.30**	0.59
	WRN-34-10			86.25	41.17	0.03
		✓		86.60	42.04	0.31
			✓	87.13	47.31	0.01
		✓	✓	**88.27**	**47.98**	0.23

Both components effectively improve robustness-accuracy trade-off

Conclusion

- We find two key issues in the invariance regularization-based AT methods.
    - (1) Gradient conflict, and (2) Mixture distribution problem
- We introduce a novel approach to achieve better robustness accuracy trade-off.



ARAT (ours)

	Defense	CIFAR10		
		Clean	AA	Sum.
ResNet-18	AT	83.77	42.42	126.19
	TRADES	81.25	48.54	129.79
	MART	82.15	47.83	129.98
	LBGAT	85.00 ± 0.47	48.85 ± 0.46	133.86 ± 0.65
	ARREST*	86.63	46.14	132.77
	AR-AT (ours)	**87.82 ± 0.19**	49.02 ± 0.47	**136.84 ± 0.33**
	AR-AT+SWA (ours)	86.44 ± 0.05	**50.28** ± 0.14	136.72 ± 0.19

ICLR
東京大学 THE UNIVERSITY OF TOKYO

Rethinking Invariance Regularization in Adversarial Training to Improve Robustness-Accuracy Trade-off

Futa Waseda [1],
Ching-Chun Chang[2],
Isao Echizen[1 2]

[1] The University of Tokyo, Japan
[2] National Institute of Information, Japan
Correspondence to: Futa Waseda <futa-waseda@g.ecc.u-tokyo.ac.jp>