A Closer Look at Accuracy vs. Robustness

Yao-Yuan Yang* Cyrus Rashtchian* Hongyang Zhang
Ruslan Salakhutdinov Kamalika Chaudhuri

University of California, San Diego
Toyota Technological Institute at Chicago
Carnegie Mellon University

July 8, 2020

* equal contribution

Adversarial example

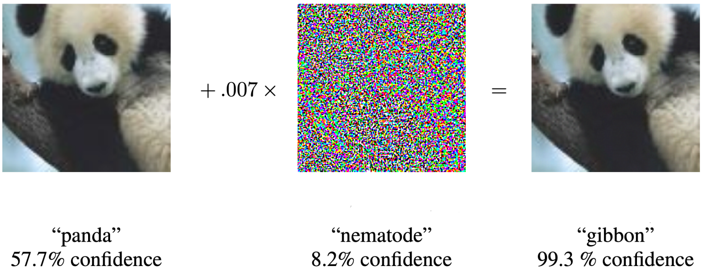


Figure: Goodfellow et al. [3]

Accuracy robustness trade-off

Trade-off between natural accuracy and adversarial accuracy is being observed on many defense algorithms (Tsipras et al. [6], Gilmer et al. [2]).

Accuracy robustness trade-off

Trade-off between natural accuracy and adversarial accuracy is being observed on many defense algorithms (Tsipras et al. [6], Gilmer et al. [2]).

Question

Is this trade-off intrinsic?

Accuracy robustness trade-off

Trade-off between natural accuracy and adversarial accuracy is being observed on many defense algorithms (Tsipras et al. [6], Gilmer et al. [2]).

Question

Is this trade-off intrinsic?

Answer

This trade-off is not intrinsic for many image classification tasks

Accuracy robustness trade-off

Trade-off between natural accuracy and adversarial accuracy is being observed on many defense algorithms (Tsipras et al. [6], Gilmer et al. [2]).

Question

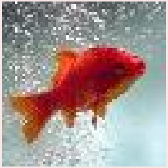
Is this trade-off intrinsic?

Answer

This trade-off is not intrinsic for many image classification tasks

What is the property that makes this trade-off not intrinsic?

r-separated: $\geq 2r$

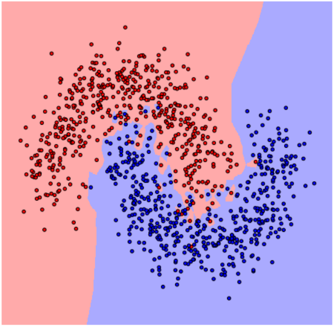


Property: r-separation

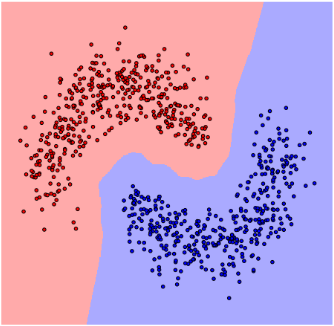
Definition (r-separation (finite sample version))

Dataset $\{\mathbf{x}_i, y_i\}_{i=1}^{N}$ is r-separated if $\forall i \neq j$:

$$y_i \neq y_j \text{ implies } \text{dist}(\mathbf{x}_i, \mathbf{x}_j) \geq 2r$$



(a) not r-separated



(b) r-separated

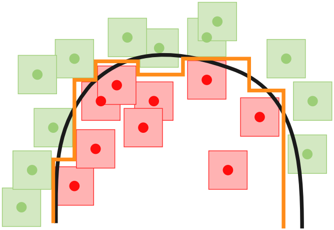
No intrinsic trade-off for r-separated data

Theorem

If data is r-separated on the support of the classes, there exists a classifier that is perfectly robust (with radius r) and accurate, based on a function with $\frac{1}{r}$-Locally Lipschitz

Definition (L-Locally Lipschitz)

A function $f : \mathcal{X} \to \mathbb{R}$ is L-Locally Lipschitz in a radius r around $\mathbf{x} \in \mathcal{X}$, if for all $\mathbf{x}'$ such that $d(\mathbf{x}, \mathbf{x}') \leq r$, it holds that: $|f(\mathbf{x}) - f(\mathbf{x}')| < L \cdot d(\mathbf{x}, \mathbf{x}')$



Real datasets are r-separated

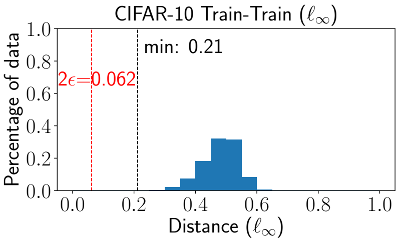
Many benchmark datasets are separated enough for common perturbation distances (ϵ) to be both robust and accurate

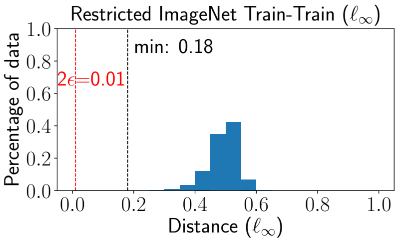
	ϵ	Required separation (2ϵ)	Train-Train separation	Test-Train separation
MNIST	0.1	0.2	0.737	0.812
CIFAR-10	0.031	0.062	0.212	0.220
SVHN	0.031	0.062	0.094	0.110
ResImageNet	0.005	0.01	0.180	0.224

Real datasets are r-separated

Many benchmark datasets are separated enough for common perturbation distances (ϵ) to be both robust and accurate

	ϵ	Required separation (2ϵ)	Train-Train separation	Test-Train separation
MNIST	0.1	0.2	0.737	0.812
CIFAR-10	0.031	0.062	0.212	0.220
SVHN	0.031	0.062	0.094	0.110
ResImageNet	0.005	0.01	0.180	0.224





Empirical observations

Comparison of different defenses

Low local Lipschitz (smooth) classifier:

- generates higher adversarial test accuracy
- increases the generalization gap

		CIFAR-10					
		train acc.	test acc.	adv test acc.	test Lipschitz	gap	adv gap
High Lip.	Natural	100.00	93.81	0.00	425.71	6.19	0.00
	GR [1]	94.90	80.74	21.32	28.53	14.16	3.94
	LLR [5]	100.00	91.44	22.05	94.68	8.56	4.50
Low Lip.	RST [7]	99.86	84.61	40.89	23.15	15.25	41.31
	AT [4]	99.84	83.51	43.51	26.23	16.33	49.94
	TRADES [8]	99.78	85.55	46.63	22.42	14.23	47.67

Conclusion

show when data distribution is r-separated, there is no intrinsic trade-off

Conclusion

show when data distribution is r-separated, there is no intrinsic trade-off

show that many real datasets are r-separated

Conclusion

show when data distribution is r-separated, there is no intrinsic trade-off

show that many real datasets are r-separated

main open question is to close the gap between robustness and accuracy

Conclusion

show when data distribution is r-separated, there is no intrinsic trade-off

show that many real datasets are r-separated

main open question is to close the gap between robustness and accuracy

- a promising direction is to reduce standard and adversarial generalization gaps

Thank you for listening.

Poster session Zoom

- Meeting ID: 927 7808 5557
- Password: 481137

More information

- Paper: https://arxiv.org/abs/2003.02460
- Code: https://git.io/JJTC6
- Blog: https://ucsdml.github.io/

Contact

- Website: http://yyyang.me/

References I

[1] Chris Finlay and Adam M Oberman. Scaleable input gradient regularization for adversarial robustness. *arXiv preprint arXiv:1905.11468*, 2019.

[2] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018.

[3] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[4] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

References II

[5] Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial robustness through local linearization. In *Advances in Neural Information Processing Systems*, pages 13847–13856, 2019.

[6] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[7] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 819–828, 2020.

[8] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019.